FINAL REPORT: EFFICIENT METHODS OF ESTIMATING THE OPERATING CHARACTERISTICS OF ITEM RESPONSE CATEGORIES AND CHALLENGE TO A NEW MODEL FOR THE MULTIPLE-CHOICE ITEM

FUMIKO SAMEJIMA

DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF TENNESSEE
KNOXVILLE, TENN. 37996-0900

NOVEMBER, 1981

Prepared under the contract number N00014-77-C-0360,
NR 150-402 with the
Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research

8

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
Final Report

2. GOVT ACCESSION NO.
AD-109 144

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Final Report: Efficient Methods of Estimating the Operating Characteristics of Item Response Categories and Challenge to a New Model for the Multiple-Choice Item

5. TYPE OF REPORT & PERIOD COVERED
Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Dr. Fumiko Samejima

8. CONTRACT OR GRANT NUMBER(s)
N00014-77-C-0360

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Psychology
University of Tennessee
Knoxville, Tennessee 37916

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
PE: 61153N; PROJ: RR 042-04
TA: RR 042-04-01
WU: NR 150-402

11. CONTROLLING OFFICE NAME AND ADDRESS
Personnel and Training Research Programs
Office of Naval Research
Arlington, Virginia 22217

12. REPORT DATE

13. NUMBER OF PAGES
243

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Operating Characteristic Estimation
Tailored Testing
Latent Trait Theory

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(Please see reverse side)

404225

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

This is the final report for the research contract N00014-77-C-0360, which started on May 1, 1977, and ended on September 30, 1981. It systematizes major findings of the research, and gives perspectives and directions of future research.

Accession For
NTIS GRA&I ☑
DTIC TAB □
Unannounced □
Justification

By
Distribution/
Availability Codes

Dist	Avail and/or Special
A	

FINAL REPORT: EFFICIENT METHODS OF ESTIMATING THE OPERATING CHARACTERISTICS OF ITEM RESPONSE CATEGORIES AND CHALLENGE TO A NEW MODEL FOR THE MULTIPLE-CHOICE ITEM

ABSTRACT

This is the final report for the research contract N00014-77-C-0360, which started on May 1, 1977, and ended on September 30, 1981. It systematizes major findings of the research, and gives perspectives and directions of future research.

This final report has been completed with the dedicated assistantship by Paul S. Changas, Charles McCarter, Shiao-Yung Chen, Zailan Bte Mohamed, and Pat Palmer.

In addition to the above five, more than twenty people have come to the principal investigator's laboratory to give her help at one time or another during the past four years and five months. The principal investigator would like to thank to all of them. Among others, Paul S. Changas, C. I. Bonnie Chen, Philip S. Livingston, Charles T. McCarter, Melanie Perkins, Robert L. Trestman and Ching-Chan Yeh have special thanks from the author.

TABLE OF CONTENTS

PREFACE

Four years and five months have passed since I started this research on May 1, 1977, and these were hectic years. During this period, so many things were designed and accomplished. Even if I am the principal investigator, I find it practically impossible to include and systematize all the important findings and implications within a single final report. I did my best within a limited amount of time, however. It is obvious that the present report should be supplemented and revised further. I plan to do so and use the result at the Advanced Seminar on Latent Trait Theory, which will be held in spring, 1982, in the vicinity of Knoxville, Tennessee, under the sponsorship of the Office of Naval Research.

There were four objectives in the original research proposal, and they can be summarized as follows.

[1] Investigation of theory and method for estimating the operating characteristics of discrete item responses, without assuming any specific mathematical forms, and without using too many examinees in the whole procedure.

[2] Investigation of the speed factor working in combination with the power factor in intellectual performance.

[3] Investigation of the random guessing behavior in testing, and the development of a new model, or new models, for the multiple-choice item.

[4] Investigation of efficient methods of estimating the ability distribution for any specific group of examinees.

Out of these four objectives, Objective [1], together with Objective [4], was very intensively pursued. The highest productivity belongs to this part of the research. Objective [3] was also successfully pursued. It provided us with valuable future perspectives and directions of research. In contrast to these three, Objective [2] was more or less dropped. To compensate for it, however, there were several other topics pursued, such as a new mathematical model for the binary item called Constant Information Model, the method of moments as the least squares solution for fitting a polynomial, Bayesian estimation of ability, and alternative estimators for the maximum likelihood estimator for the two extreme response patterns. All of these additional topics are related to the proposed objectives, but they also have the values of their own.

Recently, some researchers have started using the title, Item Response Theory, instead of Latent Trait Theory, the former of which, I believe, was first proposed by Dr. Frederic M. Lord. Although I have a great deal of respect for Dr. Lord for his long, brilliant career as a researcher and scholar, I prefer Latent Trait Theory. One of the reasons for my preference is that I see no reason why it should be changed, after so many years of presentations and publications of papers under the title of Latent Trait Theory, which include my own paper presented at the Fifth International Symposium on Multivariate Analysis, and published in Multivariate Analysis V (Krishnaiah, Ed., 1978) as a chapter. I feel that the change of the title would cause more confusion than anything else, not only among psychologists but also among mathematicians and mathematical statisticians who have become familiar with the Theory. Secondly, the term, Item Response Theory, has been used mainly by researchers whose interest is in the three-parameter logistic model in the uni-dimensional latent space. For the type of research such as mine, which covers broader areas and even includes the multi-dimensional latent space, Latent Trait Theory sounds more appropriate. In the present report, therefore, Latent Trait Theory is exclusively used for the general title, instead of Item Response Theory.

September 30, 1981

Author

I *General Background*

Latent Trait Theory can be traced back to the nineteen-forties, in the work of Lawley (Lawley, 1943) and others. In the nineteen-fifties, psychometricians like Tucker and Lord developed the basic theory as a mental test theory, and, among others, Lord integrated and published it in a Psychometric Monograph (Lord, 1952). These early works by psychometricians were joined by the latent structure analysis, which had been developed by Lazarsfeld (Lazarsfeld, 1959) and others as a theory of social attitude measurement in the area of sociology, and also by the work accomplished by Rasch (Rasch, 1960) in the context of mental measurement. These pioneer works led us to a comprehensive system of the Latent Trait Theory.

The modern mental test theory thus established originally adopted the *normal ogive model* for the conditional probability of the correct answer, given ability, or the *item characteristic function*, of the dichotomously scored test item. In the nineteen-sixties, Birnbaum (Birnbaum, 1968) proposed the *logistic model*, which is an approximation to the normal ogive model with its benefit of mathematical simplicities caused by a simple sufficient statistic for the vector of binary item scores, or the *response pattern*. Birnbaum also proposed the *three-parameter logistic model* for the multiple-choice test item, which is a modification of the logistic model and is based upon the *knowledge or random guessing principle*. Samejima (Samejima, 1969) expanded the theory to include both the *nominal and graded response levels*, in addition to the *dichotomous response level*. The graded response level assumes integers, 0 through m_g (≥ 1) , for the item score, and is further classified into two cases, the homogeneous case and the heterogeneous case (Samejima, 1972). With this generalization, we needed more than a single item characteristic function for a test item, and the conditional probability, given ability, or the *operating characteristic*, of each of the discrete responses to an item was introduced. Both the normal ogive model and the logistic model were expanded for the homogeneous case of the graded response level, which provide us with ordered, unimodal operating characteristics for all the intermediate response categories.

Sufficient conditions for a model to have a unique maximum of the operating characteristic of each and every response pattern were investigated and postulated. Bock (Bock, 1972) proposed a multinomial response model, which can either be interpreted as a model on the nominal response level or as a model in the heterogeneous case of the graded response level. Samejima (Samejima, 1973) also proposed several models on the *continuous response level*, defining the *operating density characteristic* for each continuous item response, and, later (Samejima, 1974), she expanded it to the *multi-dimensional latent space*.

In contrast to the development of the theory, its applications are still far behind. For one thing, the theory has not been well understood and used by most applied researchers. Many psychologists still bury themselves in the tautology of the classical mental test theory, although it has been pointed out (Samejima, 1977) that such core concepts in classical test theory as the *reliability coefficient* and the *validity coefficient* of a test are highly irrelevant and misleading, and that the *information functions* in Latent Trait Theory provide us with a far more relevant set of information.

In the past decade, Rasch model has become increasingly popular among certain applied researchers. The development of *adaptive testing*, or tailored testing, has also made the three-parameter logistic model popular among researchers of mental measurement. The gradual popularities of these two models do not always depend upon the relevance of these models, however. Researchers tend to choose one of those models fairly arbitrarily, and because of its availability and easiness in handling rather than their scientific convictions. The worst of all, very little effort has been put upon the model validation, which is essential in any scientific research.

The orientation we aim at in the present study is quite different from the general trends described in the preceding paragraph. We consider ourselves slaves to the truth, rather than masters who can choose their models as they wish and for their own convenience. This orientation leads us to the emphasis upon the elimination of as many assumptions as possible, and upon the model validation whenever we use one. The author hopes that the present study will stimulate some of the researchers following general trends to the extent that they wish to change their ways, following harder paths to reach the productivity of truly scientific sense.

References

[1] Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores. Chapter 17-20. Reading, Mass.: Addison-Wesley, 1968.

[2] Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 1972, 29-51.

[3] Lawley, D. N. On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 1943, 61, 273-287.

[4] Lazarsfeld, P. F. Latent structure analysis. In S. Koch (Ed.), Psychology: A study of a science. Vol. 3. New York: McGraw-Hill, 1959, 476-542.

[5] Lord, F. M. A theory of test scores. Psychometric Monograph, No. 7, 1952.

[6] Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielson and Lydiche, 1960.

[7] Samejima, F. Estimation of ability using a response pattern of graded scores. Psychometrika Monograph, No. 17, 1969.

[8] Samejima, F. A general model for free-response data. Psychometrika Monograph, No. 18, 1972.

[9] Samejima, F. Homogeneous case of the continuous response level. Psychometrika, 1973, 38, 203-219.

[10] Samejima, F. Normal ogive model on the continuous response level in the multidimensional latent space. Psychometrika, 1974, 39, 111-121.

[11] Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, 233-247.

II Research Reports

There are nineteen technical reports published during the contract period. All of them, except for three, were written by the principal investigator. The three technical reports, RR-79-2, RR-80-1 and RR-81-3, were written under the coauthorship of the principal investigator, and Philip Livingston, Robert Trestman and Paul Changas, respectively. There is one Scientific Monograph published by the Tokyo Office of the Office of Naval Research in 1980. There are two papers in the proceedings of the Computerized Adaptive Testing Conference, in 1977 and in 1979, respectively. The titles of these twenty-two research reports are listed on the following pages.

In addition to them, during the contract period, the principal investigator introduced some of the products and findings of the present research in an invited paper at the Fifth International Symposium on Multivariate Analysis, which was held at the University of Pittsburgh, in 1978. The title of the paper is Latent Trait Theory and Its Applications, and was published in Multivariate Analysis V (Krishnaiah, Ed.; North-Holland, 1980).

The twenty-two research reports can roughly be categorized into seven groups, and, in the list, they are marked with different symbols accordingly. There are eleven papers which are marked with Δ . All of them concern with the estimation of the operating characteristics of discrete item responses, and the estimation of the ability distribution. The method of moments as the least squares solution for fitting a polynomial is discussed in one paper, which is marked with ∂ . There are two papers marked with Ψ , and they are concerning the new family of models for the multiple-choice test item. There is one paper with the mark ∞ , which is an empirical study concerning the multiple-choice test item, and is related with the previous two. There are three papers on the Constant Information Model, which is a new model proposed by the principal investigator, and these papers are marked with Φ in the list. There are two papers on the computerized adaptive testing, and they are marked with Ω . Partly related with these two, there are two papers

LIST OF ONR TECHNICAL REPORTS AND OTHERS

Δ (1) Samejima, F. Estimation of the operating characteristics of item response categories I: Introduction to the Two-Parameter Beta Method. RR-77-1, 1977.

Ω (2) Samejima, F. The application of graded response models. Proceedings of the 1977 Computerized Adaptive Testing Conference (D.J. Weiss, Ed.), 28-37, 1977.

Ω (3) Samejima, F. Future directions for computerized adaptive testing (panel discussion). Proceedings of the 1977 Computerized Adaptive Testing Conference (D.J. Weiss, Ed.), 430-440, 1977.

Δ (4) Samejima, F. Estimation of the operating characteristics of item response categories II: Further development of the Two-Parameter Beta Method. RR-78-1, 1978.

Δ (5) Samejima, F. Estimation of the operating characteristics of item response categories III: The Normal Approach Method and the Pearson System Method. RR-78-2, 1978.

Δ (6) Samejima, F. Estimation of the operating characteristics of item response categories IV: Comparison of the different methods. RR-78-3, 1978.

Δ (7) Samejima, F. Estimation of the operating characteristics of item response categories V: Weighted Sum Procedure in the Conditional P.D.F. Approach. RR-78-4, 1978.

Δ (8) Samejima, F. Estimation of the operating characteristics of item response categories VI: Proportioned Sum Procedure in the Conditional P.D.F. Approach. RR-78-5, 1978.

Δ (9) Samejima, F. Estimation of the operating characteristics of item response categories VII: Bivariate P.D.F. Approach with Normal Approach Method. RR-78-6, 1978.

Φ (10) Samejima, F. Constant Information Model: A new, promising item characteristic function. RR-79-1, 1979.

∂ (11) Samejima, F. and P. S. Livingston. Method of moments as the least squares solution for fitting a polynomial. RR-79-2, 1979.

Φ (12) Samejima, F. Convergence of the conditional distribution of the maximum likelihood estimate, given latent trait, to the asymptotic normality: Observations made through the Constant Information Model. RR-79-3, 1979.

Ψ (13) Samejima, F. A new family of models for the multiple-choice item. RR-79-4, 1979.

Φ (14) Samejima, F. Constant Information Model on the dichotomous response level. Proceedings of the 1979 Computerized Adaptive Testing Conference (D.J. Weiss, Ed.), 145-163, 1979.

Ψ (15) Samejima, F. Research on the multiple-choice test item in Japan: Toward the validation of mathematical models. ONR-Tokyo, Scientific Monograph 3, April, 1980.

∞ (16) Samejima, F. & R. L. Trestman. Analysis of Iowa data I: Initial study and findings. RR-80-1, 1980.

Δ (17) Samejima, F. Estimation of the operating characteristics when the test information of the Old Test is not constant I: Rationale. RR-80-2, 1980.

¶ (18) Samejima, F. Is Bayesian estimation proper for estimating the individual's ability? RR-80-3, 1980.

Δ (19) Samejima, F. Estimation of the operating characteristics when the test information of the Old Test is not constant II: Simple Sum Procedure of the Conditional P.D.F. Approach/ Normal Approach Method using three subtests of the Old Test. RR-80-4, 1980.

¶ (20) Samejima, F. An alternative estimator for the maximum likelihood estimator for the two extreme response patterns. RR-81-1, 1981.

Δ (21) Samejima, F. Estimation of the operating characteristics when the test information of the Old Test is not constant II: Simple Sum Procedure of the Conditional P.D.F. Approach/Normal Approach Method using three subtests of the Old Test. No. 2. RR-81-2, 1981.

Δ (22) Samejima, F. & P. S. Changas. How small the number of the test items can be for the basis of estimating the operating characteristics of the discrete responses to unknown test items. RR-81-3, 1981.

concerning Bayesian vs. the maximum likelihood estimation of ability, which is marked with ¶ .

The contents and the main findings of these papers will be integrated and summarized in the following chapters. The reader will also find out how these seemingly separate topics are related, and how we can use them together to accomplish useful research.

III Estimation of the Operating Characteristics of the Discrete Item Responses and That of Ability Distributions: I

As we have seen in the preceding chapter, there are eleven papers written on these two subjects, and one paper on the method of moments which takes an important role in the methods and approaches for these estimations. In the present chapter, we shall start integrating the rationale, data and methods of this part of the research, and organize them into several sections.

(III.1) Relationship between the Estimation of the Operating Characteristics and that of Ability Distributions

By discrete item responses we mean any discrete answer to the item, including both free responses and multiple-choice responses. When free responses are treated as they are, or more or less categorized depending upon their mutual similarities, they provide us with nominal responses. If we use a dichotomous scoring stategy by categorizing them into two categories, i.e., "correct" and "incorrect", then they will be treated as dichotomous responses. If we adopt a more graded scoring strategy by categorizing them into more than two categories, i.e., 0 through m_g for item g, depending upon their closeness to the correct answer, then they will be treated as graded responses. In each case, we have discrete item responses.

Let θ be ability, or latent trait, which assumes any real number. Let $f(\theta)$ be the density function of ability θ for a given group of examinees. We denote the set of all the discrete responses to item g by K_g, and its element by k_g or h_g. Then the density function, $f(\theta)$, can be written as

$$(3.1) \qquad f(\theta) = \sum_{k_g \varepsilon K_g} f_{k_g}(\theta)\, p(k_g) \,,$$

where $f_{k_g}(\theta)$ is the density function of ability θ for the subgroup of examinees whose responses to item g are uniformly k_g, and $p(k_g)$ is the probability assigned to the subgroup within the total group of examinees. We can write for the operating characteristic, $P_{k_g}(\theta)$, of the discrete item response k_g such that

$$(3.2) \qquad P_{k_g}(\theta) = f_{k_g}(\theta)\, p(k_g) \left[\sum_{h_g \varepsilon K_g} f_{h_g}(\theta)\, p(h_g) \right]^{-1} .$$

Equation (3.2) indicates that the estimated operating characteristic of a discrete item response k_g can be obtained by the ratio of its estimated absolute frequency of ability to the absolute frequency for the whole set, K_g . Throughout the present study, this ratio is the estimated operating characteristic we adopt. Any method for estimating the operating characteristics of discrete item responses includes, therefore, the estimation of two or more ability distributions. In other words, those methods and approaches developed in the present study are not only for the estimation of the operating characteristics but also for the estimation of ability distributions.

There is a certain invariance property in the estimated operating characteristic over the transformation of the latent trait, which is not shared by the estimated probability density of ability. Let τ be a strictly increasing and differentiable function of θ . We have for the densities, $f^*(\tau)$ and $f^*_{k_g}(\tau)$, for the transformed latent trait τ , such that

$$(3.3) \qquad f^*(\tau) = f(\theta) \frac{d\theta}{d\tau} ,$$

and

$$(3.4) \qquad f^*_{k_g}(\tau) = f_{k_g}(\theta) \frac{d\theta}{d\tau} ,$$

for any discrete response $k_g \varepsilon K_g$. From (3.2) and (3.4) it is obvious that for the operating characteristic, $P^*_{k_g}(\tau)$, we have

$$(3.5) \qquad P^*_{k_g}(\tau) = P_{k_g}(\theta) ,$$

which indicates the invariance of the estimated operating characteristic over the transformation of the latent trait.

(III.2) No Mathematical Forms Are Assumed for the Operating Characteristics of the Unknown Test Items

Most researchers preassume some mathematical model for the operating characteristics of the item responses of their unknown test items. In such a case, the estimation of the operating characteristics is converted to the estimation of a small number of item parameters. This simplification will make it easy for us to conduct our research. On the other hand, in so doing, we may distort the psychological reality, which is the very object of our research, by molding it into some irrelevant model. Thus both the deductive and inductive validations of the model are by far the most important when we adopt any mathematical model. In other words, the model must follow a rationale which also explains the psychological reality behind our data, and, once they were analyzed, we must validate the model by finding out if the internal consistency exists.

The importance of the model validation seems to be forgotten by many researchers, however. To give an example, the popularity of Rasch model mainly depends upon its mathematical simplicity, which comes from the fact that it has only one parameter, i.e., the difficulty. Very few researchers stop to think, however, whether this particular model and its simplicity are appropriate for their data, nor do they try to find out the validity of the model by checking the internal consistency in their results. Anocher example is the way many researchers use the three-parameter logistic model for their data of multiple-choice test items. The rationale behind the model is the knowledge or random guessing principle, which is rather unlikely to be the case in most multiple-choice testing situations. Among others, the fact that they are ready to accept a value which is less than the reciprocal of the number of the alternatives of a specified multiple-choice test item as the third parameter, i.e., the guessing parameter, is nothing but defeating itself.

To avoid the possibility of adopting an irrelevant mathematical model, the best solution will be to develop methods of estimating the operating characteristics of the discrete item responses without assuming any mathematical forms. In the present study, this direct approach to the operating characteristics is consistently used. Although it creates more difficulty and requires more labors in developing our methods and approaches, it is worth our effort considering the due cause we have. The reader will find similar attempts in the works by Lord (Lord, 1970) and Levine (Levine, 1980), i.e., estimation of the operating characteristics without assuming any mathematical forms.

(III.3) Small Numbei of Examinees in the Calibration Data

For a relatively few researchers whose calibration data are obtained from institutes like Educational Testing Service, it is easy to use those which were collected upon several hundred thousand examinees. For most researchers who do their research in university environments, however, the situation is quite different. It may be extremely difficult for them to find even one thousand volunteer students for their subjects. For this reason, it is necessary that we should investigate and develop methods of estimating the operating characteristics which do not require more than several hundred examinees for our calibration data.

This is one of the important considerations in the present study. Our calibration data are based upon five hundred hypothetical examinees, whose ability levels are at one hundred equally spaced positions on the ability dimension, with five examinees being placed at each position. This configuration can be considered as an approximation to a uniform distribution of ability. To be specific, the five hundred ability levels range from -2.475 to 2.475, with the equal steps of 0.05. The uniform distribution has, therefore, the density of 0.2, for the interval of ability θ, $(-2.5, 2.5)$, as is shown in Figure 3-3-1.

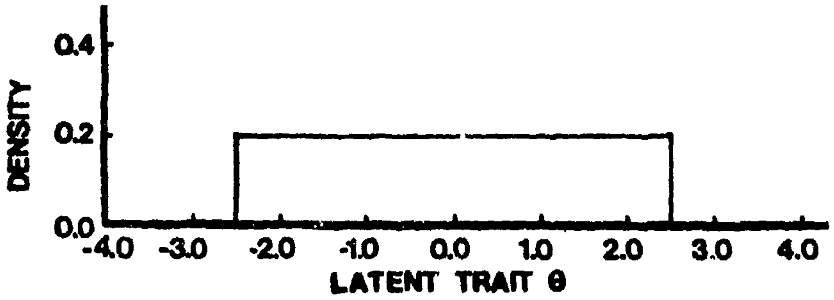


FIGURE 3-3-1

Ability Distribution of Our Hypothetical Examinees. Actually, the Five Hundred Examinees Are Placed at the One Hundred Equally Spaced Positions from -2.475 to 2.475 , with Five Examinees Sharing Each Position.

(III.4) Old Test

It is assumed that there exists a set of test items whose operating characteristics are known, and our examinees have taken the test, as well as a set of test items whose operating characteristics are to be estimated. We call the first set of test items Old Test, and the estimation of the operating characteristics of the test items of the second set is based upon the examinees' performances on the Old Test.

The methods and approaches developed on this assumption are directly useful in such a situation that, in adaptive testing, we have a well-constructed item pool, but we want to add more test items to our item pool. Another suitable situation will be that we have a relatively small number of well developed test items which have a high content validity for our purpose of measurement, and on the trial-and-error basis we have obtained confirmed mathematical model or models for separate test items with respect to their deductive and inductive validities, so that we shall be able to use them as our Old Test.

This assumption of the existence of the Old Test is a restriction, which we may wish to eliminate so that we shall be able to expand the applicability of our methods and approaches to the situation where we must start the calibration of the operating characteristics from scratch. There are two different attempts for this purpose, which will be discussed in a later chapter.

In the present study, a set of thirty-five test items has been chosen as our original Old Test. Each of these thirty-five items has three graded item score categories, and follows the normal ogive model such that

$$(3.6) \qquad P_{x_g}(\theta) = [2\pi]^{-1/2} \int_{a_g(\theta-b_{x_g+1})}^{a_g(\theta-b_{x_g})} e^{-u^2/2}\,du \quad ,$$

where x_g $(=0,1,\ldots,m_g)$ is the graded item score of item g , $P_{x_g}(\theta)$ is its operating characteristic, a_g (> 0) is the item discrimination parameter, and b_{x_g} is the item response difficulty parameter which satisfies

$$(3.7) \qquad -\infty = b_0 < b_1 < \ldots < b_{m_g} < b_{m_g+1} = \infty \ .$$

The item parameters and item response parameters of these thirty-five test items are shown in Table 3-4-1 . We have also used nine different subtests of the original Old Test as our Old Test on different occasions, and these subtests are shown in the same table by indicating the test items by crosses. The numbers of test items in these subtests range from five to twenty-five.

We can write for the item response information function, $I_{x_g}(\theta)$, such that

$$(3.8) \qquad I_{x_g}(\theta) = -\frac{\partial^2}{\partial\theta^2}\log P_{x_g}(\theta) \ ,$$

and the item information function, $I_g(\theta)$, is given as the conditional

TABLE 3-4-1

Item Parameters of the Test Items of Our Old Tests.

Item g	a_g	b_1	b_2	Subtests								
				1	2	3	4	5	6	7	8	9
1	1.8	-4.75	-3.75		x							
2	1.9	-4.50	-3.50		x					x		
3	2.0	-4.25	-3.25		x				x		x	
4	1.5	-4.00	-3.00		x			x				x
5	1.6	-3.75	-2.75		x							
6	1.4	-3.50	-2.50	x	x		x	x	x	x		
7	1.9	-3.00	-2.00	x	x		x					
8	1.8	-3.00	-2.00	x	x		x	x			x	
9	1.6	-2.75	-1.75	x	x		x		x			
10	2.0	-2.50	-1.50	x	x		x	x		x		
11	1.5	-2.25	-1.25	x	x	x	x					x
12	1.7	-2.00	-1.00	x	x	x	x	x	x			
13	1.5	-1.75	-0.75	x		x					x	
14	1.4	-1.50	-0.50	x		x		x		x		
15	2.0	-1.25	-0.25	x		x			x			
16	1.6	-1.00	0.00	x		x		x				
17	1.8	-0.75	0.25	x		x						
18	1.7	-0.50	0.50	x		x		x	x	x	x	x
19	1.9	-0.25	0.75	x		x						
20	1.7	0.00	1.00	x		x		x				
21	1.5	0.25	1.25	x		x			x			
22	1.8	0.50	1.50	x		x		x		x		
23	1.4	0.75	1.75	x	x	x	x				x	
24	1.9	1.00	2.00	x	x	x	x	x	x			
25	2.0	1.25	2.25	x	x	x	x					x
26	1.6	1.50	2.50	x	x		x	x		x		
27	1.7	1.75	2.75	x	x		x		x			
28	1.4	2.00	3.00	x	x		x	x			x	
29	1.9	2.25	3.25	x	x		x					
30	1.6	2.50	3.50	x	x		x	x	x	x		
31	1.5	2.75	3.75		x							
32	1.7	3.00	4.00		x			x				x
33	1.8	3.25	4.25		x				x		x	
34	2.0	3.50	4.50		x					x		
35	1.4	3.75	4.75		x							

expectation of the item response information function, i.e.,

$$(3.9) \qquad I_g(\theta) = \sum_{x_g=0}^{m_g} I_{x_g}(\theta)\, P_{x_g}(\theta) .$$

The response pattern of the set of n test items is the set of the n item scores such that

$$(3.10) \qquad V = (x_1, x_2, \ldots, x_g, \ldots, x_n)' .$$

By virtue of the local independence (Lord and Novick, 1968, Chapter 16), the operating cnaracteristic of the response pattern V is given as the product of the n operating characteristics of the item scores, so that we have

(3.11) $$P_V(\theta) = \prod_{x_g \in V} P_{x_g}(\theta) \ .$$

We can write for the response pattern information function, $I_V(\theta)$, such that

(3.12) $$I_V(\theta) = -\frac{\partial^2}{\partial\theta^2}\log P_V(\theta) = \sum_{x_g \in V} I_{x_g}(\theta) \ ,$$

and the test information function, $I(\theta)$, is the conditional expectation of the response pattern information functions. We obtain

(3.13) $$I(\theta) = \sum_V I_V(\theta)\, P_V(\theta) = \sum_{g=1}^{n} I_g(\theta) \ .$$

The square root of the test information function of the Old Test has an important role in the present study, which will be described in later sections. For the original Old Test, this function of θ is approximately constant (= 4.65) for the interval of θ of our interest, i.e., approximately, (-3.0, 3.0) . For the nine subtests of the original Old Test, this function is not constant, but is either a unimodal or a bimodal function of θ . The square root of the test information function for each of the ten Old Tests which were used in the present study is shown in Figure 3-4-1.

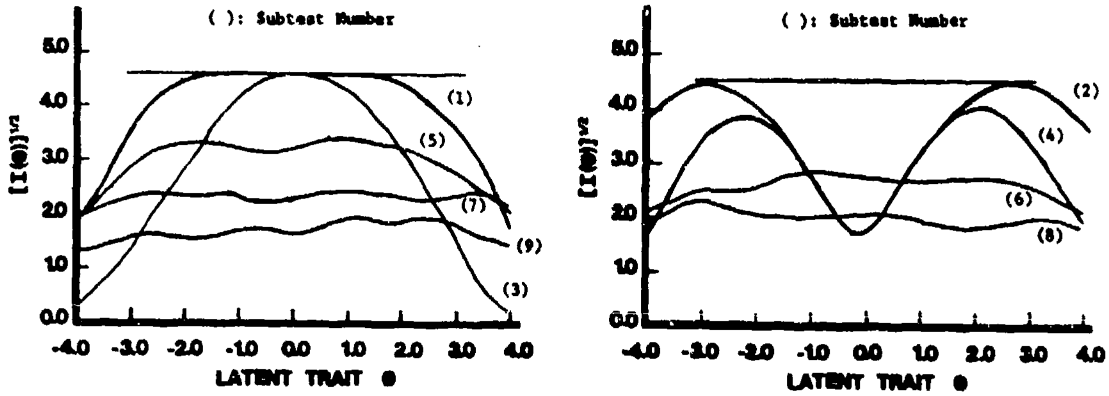


FIGURE 3-4-1

Square Root of the Test Information Function of Each of Subtests 1 through 9 , Together with the One for the Original Old Test (≑4.65).

(III.5) Set of Five Hundred Maximum Likelihood Estimates

The maximum likelihood estimate of the examinee's ability when each of the n items follows the normal ogive model can be obtained numerically (Samejima, 1969, 1972), by using the operating characteristic $P_V(\theta)$ as the likelihood function. Let $A_{x_g}(\theta)$ be the basic function of item score x_g , which is defined by

$$(3.14) \qquad A_{x_g}(\theta) = \frac{\partial}{\partial\theta} \log P_{x_g}(\theta) \ .$$

We can write for the maximum likelihood estimate $\hat{\theta}_V$ for the response pattern V such that

$$(3.15) \qquad \sum_{x_g \varepsilon V} A_{x_g}(\hat{\theta}_V) = 0 \ .$$

In the normal ogive model, this basic function is a strictly decreasing function of θ , and the two asymptotes of the basic function are 0 and $-\infty$ for the lowest extreme response pattern $(0,0,\ldots,0)$, ∞ and 0 for the highest extreme response pattern, $(m_1,m_2,\ldots,m_n)$, and $-\infty$ and ∞ for all the other intermediate response patterns.

In our study, by the Monte Carlo method, we calibrated, for each hypothetical examinee, the response pattern of the n test items of the Old Test, and based upon this response pattern the maximum likelihood estimate of his ability was obtained. This set of five hundred maximum likelihood estimates takes an essential role in the calibration of the operating characteristics of each of our unknown test items.

The maximum likelihood estimate has such an asyptotic property that the estimate is conditionally unbiased and normally distributed with θ and $[I(\theta)]^{-1/2}$ as its two parameters, given θ . It has been observed (Samejima, 1975, 1977a, 1977b) that this asymptotic normal distribution can be used as a good approximation to the conditional distribution of $\hat{\theta}_V$, given θ , even when the number of test items is not so large and the amount of test information is relatively small. Throughout the present study, this approximation is effectively used.

(III.6) Unknown Test Items Whose Operating Characteristics Are to Be Estimated

There are ten hypothetical, binary test items, and throughout the present study, our target is the estimation of the operating characteristic of $x_g = 1$, or the item characteristic function, for each of these ten binary items. Let $P_h(\theta)$ be the item characteristic function of the unknown test item h . For each item, this item characteristic function follows the normal ogive model, such that

$$P_h(\theta) = [2\pi]^{-1/2} \int_{-\infty}^{a_h(\theta-b_h)} e^{-u^2/2}\, du \quad (3.16)$$

The discrimination parameter a_h and the difficulty parameter b_h are shown in Table 3-6-1 for each of these ten binary test items.

TABLE 3-6-1

Item Discrimination Parameter a_h and Item Difficulty Parameter b_h of Each of Ten Binary Items

Item h	a_h	b_h
1	1.5	-2.5
2	1.0	-2.0
3	2.5	-1.5
4	1.0	-1.0
5	1.5	-0.5
6	1.0	0.0
7	2.0	0.5
8	1.0	1.0
9	2.0	1.5
10	1.0	2.0

There is no doubt for the necessity of using more varieties of operating characteristics for the unknown test items, including unimodal functions, functions with non-zero asymptotes, and so on. Because of the amount of work done in the present study, however, this has to wait for future research. The author hopes that some other researchers will get interested in conducting such research, using the methods and approaches developed in the present study.

(III.7) Use of Robust, Indirect Information

Lord adopted his own method (Lord, 1969) of estimating true score distributions from the observed score distributions in his attempt (Lord, 1970) of estimating the item characteristic functions of the SAT Verbal Test items without preassuming any mathematical forms. He excluded the item under study from the total test in defining the test score. This direct approach to the operating characteristics does not require Old Test, and we can start from the direct observation of the sample test score distributions. The number of examinees Lord used in his calibration of the item characteristic functions is 103,275 . This valuable study by Lord provides us with a methodology which we can use for empirical data which are found in large institutes like Educational Testing Service.

There is no question that a large sample size is desirable in the estimation of the operating characteristics. There is a necessity, however, that we should develop methodologies which are applicable for much smaller groups of examinees. Levine (Levine, 1980) developed a method with this consideration in mind. Following the present study by the author, he used Old Test as the basis of calibrating the operating characteristics of unknown test items. In his method, Levine introduced a set of orthonormal eigenfunctions, the number of which does not exceed the number of all possible response patterns of the Old Test. In practice, this number is much less than this maximal value, and it is interesting to note that it depends not only upon the number of test items in the Old Test but also upon the number of examinees. In other words, Levine's method involves a certain trade-off relationship between the number of examinees and that of test items in the Old Test. He has tried his own method using the author's simulated data based upon the original Old Test (cf. Section III.4) and the five hundred hypothetical examinees (cf. Section III.3), and his results turned out to be successful. He also tried his method using SAT test items (Levine, 1981), using somewhat larger numbers of examinees, like one thousand.

In using a small number of examinees as the basis of the calibration of operating characteristics, we need some additional information other than the one which is directly observable, such as the observed test score distribution, the response pattern, and so on. Such indirect information must be robust to the fluctuation caused by a small sample size. In the present study, the conditonal moments of θ , given its maximum likelihood estimate $\hat{\theta}$, serves for the purpose. In other words, instead of approaching the ability distribution directly as is the case with Lord's method and Levine's method, we focus our attention to the conditional distribution of ability θ , given its maximum likelihood estimate $\hat{\theta}$, or the bivariate distribution of θ and $\hat{\theta}$. Thus the estimated unconditional ability distribution is obtained as an aggregate of the estimated conditional density function of θ , given $\hat{\theta}$, or in the form of integration of the estimated bivariate density function of θ and $\hat{\theta}$.

Let us assume that the square root of the test information function of our Old Test is constant for the interval of θ of our interest, as is the case with our original Old Test. We shall denote the conditional density of $\hat{\theta}$, given ability θ , by $\psi(\hat{\theta}|\theta)$. By virtue of the asympototic normality of the conditional distribution of $\hat{\theta}$, given θ , $\psi(\hat{\theta}|\theta)$ is approximated by the normal density function, with θ and $[I(\theta)]^{-1/2}$ as its parameters. Let σ denote the constant value of $[I(\theta)]^{-1/2}$. The first through fourth derivatives of $\psi(\hat{\theta}|\theta)$ with respect to θ can be written as follows.

(3.17) $$\frac{\partial}{\partial\hat{\theta}} \psi(\hat{\theta}|\theta) = -\psi(\hat{\theta}|\theta)\sigma^{-2}(\hat{\theta}-\theta) .$$

(3.18) $$\frac{\partial^2}{\partial\hat{\theta}^2} \psi(\hat{\theta}|\theta) = \psi(\hat{\theta}|\theta)\sigma^{-2}[\sigma^{-2}(\hat{\theta}-\theta)^2 - 1] .$$

(3.19) $$\frac{\partial^3}{\partial\hat{\theta}^3} \psi(\hat{\theta}|\theta) = 3\psi(\hat{\theta}|\theta)\sigma^{-4}(\hat{\theta}-\theta) - \psi(\hat{\theta}|\theta)\sigma^{-6}(\hat{\theta}-\theta)^3 .$$

(3.20) $$\frac{\partial^4}{\partial\hat{\theta}^4} \psi(\hat{\theta}|\theta) = 3\psi(\hat{\theta}|\theta)\sigma^{-4} - 6\psi(\hat{\theta}|\theta)\sigma^{-6}(\hat{\theta}-\theta)^2 + \psi(\hat{\theta}|\theta)\sigma^{-8}(\hat{\theta}-\theta)^4 .$$

Let $g(\hat{\theta})$ be the density function of the maximum likelihood estimate $\hat{\theta}$. We can write

(3.21) $$g(\hat{\theta}) = \int_{-\infty}^{\infty} \psi(\hat{\theta}|\theta)f(\theta)\, d\theta .$$

Let us assume that this density function, $g(\hat{\theta})$, is four times differentiable. We obtain for the conditional expectation of θ , given $\hat{\theta}$, and the second, third and fourth conditional moments of θ about the mean, given $\hat{\theta}$,

(3.22) $$E(\theta|\hat{\theta}) = \hat{\theta} + \sigma^2 \frac{d}{d\hat{\theta}} \log g(\hat{\theta}) = \hat{\theta} + \sigma^2[\frac{d}{d\hat{\theta}} g(\hat{\theta})][g(\hat{\theta})]^{-1} .$$

(3.23) $$\begin{aligned} \text{Var.}(\theta|\hat{\theta}) &= \sigma^2[1 + \sigma^2 \frac{d^2}{d\hat{\theta}^2} \log g(\hat{\theta})] \\ &= \sigma^2[1 + \sigma^2\{ \frac{d^2}{d\hat{\theta}^2} g(\hat{\theta})\cdot g(\hat{\theta}) - [\frac{d}{d\hat{\theta}} g(\hat{\theta})]^2\}\{g(\hat{\theta})\}^{-2}] . \end{aligned}$$

(3.24) $$E[\{\theta-E(\theta|\hat{\theta})\}^3|\hat{\theta}] = \sigma^6[\frac{d^3}{d\hat{\theta}^3} \log g(\hat{\theta})] .$$

and

(3.25) $$\begin{aligned} E[\{\theta-E(\theta|\hat{\theta})\}^4|\theta] = \sigma^4[3 &+ 6\sigma^2\{ \frac{d^2}{d\hat{\theta}^2} \log g(\hat{\theta})\} \\ &+ 3\sigma^4\{ \frac{d^2}{d\hat{\theta}^2} \log g(\hat{\theta})\}^2 + \sigma^4\{ \frac{d^4}{d\hat{\theta}^4} \log g(\hat{\theta})\}] . \end{aligned}$$

We can see from the above four formulas that these conditional moments are specified exclusively by $\hat{\theta}$, $g(\hat{\theta})$ and σ . Note, moreover, that if the density function, $g(\hat{\theta})$, is estimated, then these conditional moments are obtainable for any value of $\hat{\theta}$ within its meaningful interval. The first through fourth derivatives of $\log g(\hat{\theta})$ can be written as follows.

(3.26) $$\frac{d}{d\hat{\theta}} \log g(\hat{\theta}) = \frac{d}{d\hat{\theta}} g(\hat{\theta}) [g(\hat{\theta})]^{-1} .$$

(3.27) $$\frac{d^2}{d\hat{\theta}^2} \log g(\hat{\theta}) = [g(\hat{\theta}) \cdot \frac{d^2}{d\hat{\theta}^2} g(\hat{\theta}) - \{\frac{d}{d\hat{\theta}} g(\hat{\theta})\}^2][g(\hat{\theta})]^{-2} .$$

(3.28) $$\frac{d^3}{d\hat{\theta}^3} \log g(\hat{\theta}) = [\{g(\hat{\theta})\}^2 \cdot \frac{d^3}{d\hat{\theta}^3} g(\hat{\theta}) - 3g(\hat{\theta}) \cdot \frac{d}{d\hat{\theta}} g(\hat{\theta}) \cdot \frac{d^2}{d\hat{\theta}^2} g(\hat{\theta})$$
$$+ 2\{ \frac{d}{d\hat{\theta}} g(\hat{\theta})\}^3][g(\hat{\theta})]^{-3} .$$

(3.29) $$\frac{d^4}{d\hat{\theta}^4} \log g(\hat{\theta}) = [\{g(\hat{\theta})\}^3 \cdot \frac{d^4}{d\hat{\theta}^4} g(\hat{\theta})$$
$$- 4\{g(\hat{\theta})\}^2 \cdot \frac{d}{d\hat{\theta}} g(\hat{\theta}) \cdot \frac{d^3}{d\hat{\theta}^3} g(\hat{\theta})$$
$$- 3\{g(\hat{\theta})\}^2 \{ \frac{d^2}{d\hat{\theta}^2} g(\hat{\theta})\}^2$$
$$+ 12 g(\hat{\theta}) \{ \frac{d}{d\hat{\theta}} g(\hat{\theta})\}^2 \cdot \frac{d^2}{d\hat{\theta}^2} g(\hat{\theta})$$
$$- 6\{ \frac{d}{d\hat{\theta}} g(\hat{\theta})\}^4][g(\hat{\theta})]^{-4} .$$

We notice that, since σ is obtained as the reciprocal of the square root of the test information function of the Old Test, all we need is to estimate the density function $g(\hat{\theta})$ from the set of N maximum likelihood estimates, $\hat{\theta}_s$ $(s=1,2,\ldots,N)$, with the consideration of making the resultant density function four times differentiable. This can be done by using the method of moments (Elderton and Johnson, 1969), and approximating a polynomial to the density function $g(\hat{\theta})$. The rationale behind this method will be given in Chapter 4.

(III.8) Transformation of Ability θ to τ

We notice that the relatively simple formulas, (3.22) through (3.25), for the conditional moments of ability θ, given its maximum likelihood estimate $\hat{\theta}$, are true only when the square root of the test information function is constant for the interval of ability of our interest, as is the case with our original Old Test. As we have seen earlier (cf. Section III.4), for all the other nine Old Tests, i.e., subtests of the original Old Test, the square root of the test information function is not constant. When we use one of these nine subtests as our Old Test, therefore, (3.22) through (3.25) are no longer true as they are. This problem can be solved by transforming θ, in such a way that the resultant transformed latent trait τ has a constant value for the square root of the test information function, $[I^*(\tau)]^{1/2}$, for the meaningful interval of τ.

Let τ be a function of θ, such that

$$(3.30) \qquad \tau = \tau(\theta) ,$$

which is strictly increasing in θ. The operating characteristic, $P^*_{x_g}(\tau)$, of the item response x_g defined for the transformed latent trait τ equals the original operating characteristic, $P_{x_g}(\theta)$, which is obvious from its definition as the conditional probability. Thus we can write

$$(3.31) \qquad P^*_{x_g}(\tau) = P^*_{x_g}[\tau(\theta)] = P_{x_g}(\theta) .$$

From (3.31) and (3.8), we can write for the item response information function, $I^*_{x_g}(\tau)$, such that

$$(3.32) \qquad I^*_{x_g}(\tau) = -\frac{\partial^2}{\partial\tau^2} \log P^*_{x_g}(\tau)$$

$$= I_{x_g}(\theta)\, [\frac{d\theta}{d\tau}]^2 - \frac{\partial}{\partial\theta} \log P_{x_g}(\theta) . \frac{d^2\theta}{d\tau^2} .$$

From this result, we have for the item information function $I^*_g(\tau)$,

$$(3.33)\qquad I^*_g(\tau) = \sum_{x_g=0}^{m_g} I^*_{x_g}(\tau)\, P^*_{x_g}(\tau) = I_g(\theta)\, [\frac{d\theta}{d\tau}]^2 ,$$

since

$$(3.34)\qquad \sum_{x_g=0}^{m_g} \frac{\partial}{\partial\theta} P_{x_g}(\theta) = 0 .$$

It can be seen that, with the response pattern V , we obtain similar results, such that

$$(3.35)\qquad P^*_V(\tau) = P^*_V[\tau(\theta)] = P_V(\theta)$$

for the operating characteristic, $P^*_V(\tau)$, and

$$(3.36)\qquad I^*_V(\tau) = I_V(\theta)\, [\frac{d\theta}{d\tau}]^2 - \frac{\partial}{\partial\theta} \log P_V(\theta)\, \frac{d^2\theta}{d\tau^2}$$

for the information function, $I^*_V(\tau)$. We can write for the test information function $I^*(\tau)$ either from (3.36) or from (3.33) such that

$$(3.37)\qquad I^*(\tau) = I(\theta)\, [\frac{d\theta}{d\tau}]^2 ,$$

and, since τ is a strictly increasing function of θ , we have

$$(3.38)\qquad [I^*(\tau)]^{1/2} = [I(\theta)]^{1/2} \frac{d\theta}{d\tau} .$$

Let C be an arbitrary constant for the square root of the test information function, $[I^*(\tau)]^{1/2}$. From (3.38) we can write

$$(3.39)\qquad \frac{d\tau}{d\theta} = C^{-1} [I(\theta)]^{1/2} .$$

Thus we obtain for the transformation of θ to τ such that

$$(3.40) \qquad \tau = c^{-1} \int [I(\theta)]^{1/2} \, d\theta + d \ ,$$

where d is an arbitrary constant for adjusting the origin of τ .

In practice, this transformation will be much more simplified if we approximate the function, $[I(\theta)]^{1/2}$ by a polynomial of an appropriate degree, using the method of moments. The detail of this process will be given in Chapter 5.

We can write for the density function, $f^*(\tau)$, of the transformed ability

$$(3.41) \qquad f^*(\tau) = f(\theta) \frac{d\theta}{d\tau} \ .$$

This equation indicates that the new density function thu- obtained is no longer uniform, as is the case with our density function of θ . Figure 3-7-1 illustrates two examples of $f^*(\tau)$ as the results of the transformation of θ to τ , which are based upon Subtests 1 and 2 , respectively.

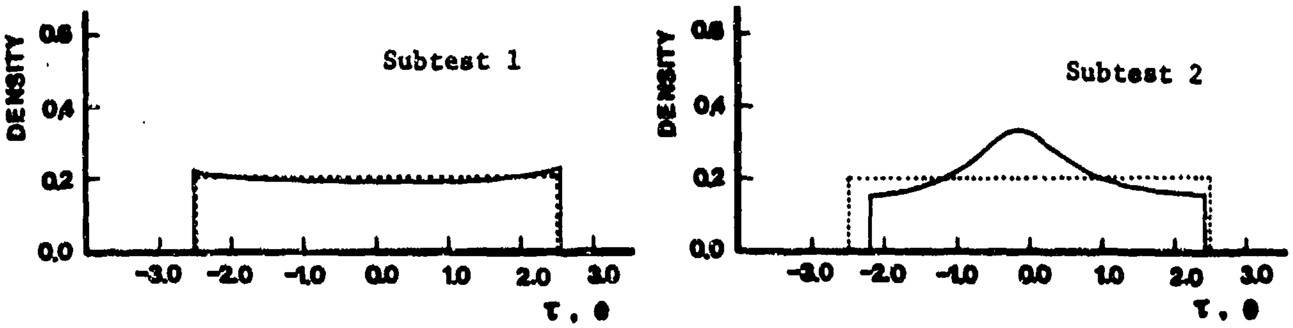


FIGURE 3-7-1

Density Function, $f^*(\tau)$, of τ Transformed from θ by the Polynomial of Degree 8 (Solid Curve), in Contrast to the Original Density Function $f(\theta)$ (Dotted Curve), when we used Subtest 1 (Left) and Subtest 2 (Right) as our Old Test, Respectively.

The maximum likelihood estimate, $\hat{\theta}$, of ability θ , which is based upon the response pattern V , can be obtained by using the operating characteristics $P_V(\theta)$ as the likelihood function.

In a similar manner, the corresponding maximum likelihood estimate, $\hat{\tau}$, can be obtained by using $P^*_V(\tau)$ as the likelihood function. By virtue of the transformation-free character of the maximum likelihood estimator, however, this second maximum likelihood estimate can also be obtained by the direct transformation of $\hat{\theta}$, such that

$$\hat{\tau} = \tau(\hat{\theta}) \tag{3.42}$$

(cf. Samejima, 1969).

REFERENCES

[1] Elderton, W. P. and N. L. Johnson. Systems of frequency curves. Cambridge University Press, 1969.

[2] Levine, M. Appropriateness measurement and the formula-score method: overview, intercorrelations and interpretations. Paper presented at the ONR Conference on Model-Based Psychological Measurement, 1980, Iowa City, Iowa.

[3] Levine, M. Ability distribution measurement for short and complex tests. Paper presented at the ONR Conference on Model-Based Psychological Measurement, 1981, Millington, Tennessee.

[4] Lord, F. M. Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). Psychometrika, 1969, 34, 259-299.

[5] Lord, F. M. Item characteristic curves estimated without knowledge of their mathematical form--a confrontation of Birnbaum's logistic model. Psychometrika, 1970, 35, 43-50.

[6] Lord, F. M. and M. R. Novick. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

[7] Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, No. 17 1969.

[8] Samejima, F. A general model for free-response data. Psychometrika Monograph, No. 18, 1972.

[9] Samejima, F. Graded response model of the latent trait theory and tailored testing. Proceedings of the First Conference on Computerized Adaptive Testing, 1975, Civil Service Commission and Office of Naval Research, 1975, pages 5-17.

[10] Samejima, F. Effects of individual optimization in setting the boundaries of dichotomous items on the accuracy of estimation. Applied Psychological Measurement, 1977(a), 1, 77-94.

[11] Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977(b), 1, 233-247.

IV Method of Moments As the Least Squares Solution for Fitting a Polynomial

The method of moments (Elderton and Johnson, 1969) was frequently used in the present study, and on many occasions it took an important role. In some situations, we fitted Pearson type density functions, and in many other situations we used polynomials. It should be noted that, when we adopt a polynomial to approximate a density function, there is a possibility that, for some range of the variable, the estimated density turns out to be negative. In practice, however, it seldom happened, and, even when it did, it did not seriously affect the process or the result of our estimation. Since the polynomial is less restrictive in its shape than many other functions which have the same number of parameters, and in addition, its derivatives are given as even simpler polynomials, the method of moments for fitting a polynomial looks promising.

In this chapter, the rationale and reason behind the success of using polynomials as functions for us to fit by the method of moments are described, and some observations are made. This part of the present final report is mainly cited from the research report RR-79-2, which includes the fine effort by one of the author's assistants, Philip Livingston.

(IV.1) Approximation to the Density Function from a Set of Observations

The method of moments was originally developed to graduate the observed frequency distribution by assuming some specific mathematical function and fitting the observed moments of up to a specified degree. This can readily be expanded to the case in which we wish to estimate a density function from a set of observations, rather than a frequency distribution.

Let μ_2 , μ_3 and μ_4 denote the second, third and fourth moments about mean of some distribution. If we preassume that the distribution should belong to the Pearson's System, then the criterion κ , which is defined by

$$(4.1) \qquad \kappa = \beta_1(\beta_2+3)^2[4(2\beta_2-3\beta_1-6)(4\beta_2-3\beta_1)]^{-1} ,$$

where β_1 and β_2 are obtained as the ratios such that

$$(4.2) \qquad \beta_1 = \mu_3^2 \, \mu_2^{-3}$$

and

$$(4.3) \qquad \beta_2 = \mu_4 \, \mu_2^{-2} ,$$

takes an important role. Substituting the sample moments for μ_2, μ_3 and μ_4 in (4.2) and (4.3), and through (4.1) we can evaluate Pearson's criterion κ, and, according to its value, we decide which type of the Pearson's system our distribution belongs to. If, for instance, κ turned out to be negative and finite, then the distribution will be of Pearson's Type I; if it turned out to be such that $\kappa = 0$, $\beta_1 = 0$ and $\beta_2 < 3$, then our distribution will be of Pearson's Type II; and so on.

Figure 4-1-1 shows the set of five hundred maximum likelihood estimates, $\hat{\theta}_s$, which was introduced in Section III.5 of the preceding chapter, in the summarized form of frequency distribution. In the same figure, also presented by a dotted line is the theoretical frequency of the maximum likelihood estimate $\hat{\theta}$, which was obtained from (3.21), using the uniform density (cf. Section III.3) for $f(\theta)$ and $n(\theta,\sigma)$ for $\psi(\hat{\theta}|\theta)$. It turned out that Pearson's criterion κ and the values of β_1 and β_2 indicated that our distribution belongs to Type II, and the frequency function obtained by the method of moments is drawn by a solid line in Figure 4-1-1.

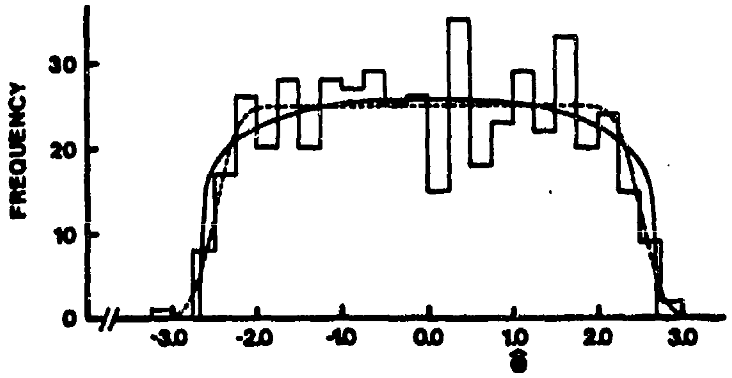


FIGURE 4-1-1

Frequency Distribution of the Five Hundred Maximum Likelihood Estimates (Histogram), Pearson's Type II Frequency Function Fitted by the Method of Moments (Solid Curve) and the Theoretical Frequency Function of the Maximum Likelihood Estimate, $\hat{\theta}$.

In contrast to this result, Figure 4-1-2 presents similar results, which were obtained by approximating the frequency function by

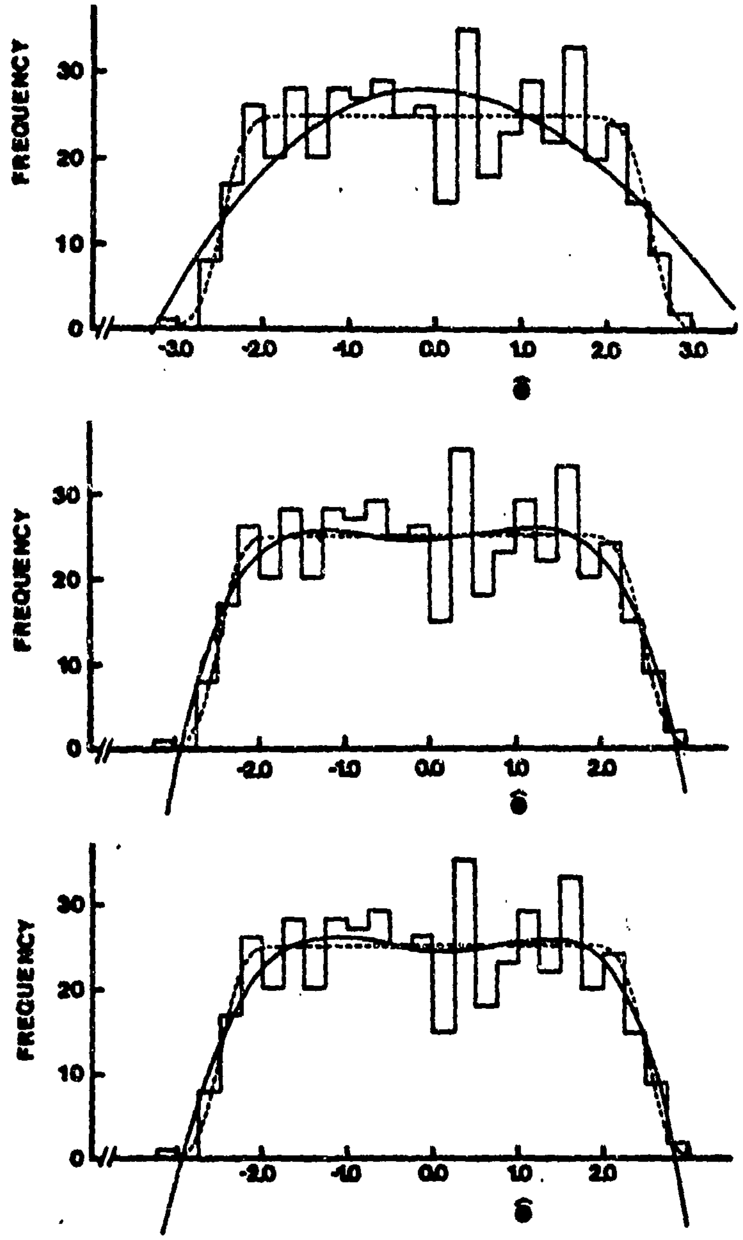


FIGURE 4-1-2

Frequency Distribution of the Five Hundred Maximum Likelihood Estimates (Histogram), the Polynomial Fitted by the Method of Moments (Solid Curve and the Theoretical Frequency Function of the Maximum Likelihood Estimate, $\hat{\delta}$. The Three Polynomials are of Degrees 3 , 4 and 5 , Respectively.

the polynomials of degrees 3, 4 and 5 using the method of moments, respectively. Comparison of these results with Figure 4-1-1 may make us prefer polynomials to Pearson type frequency functions, because of their flexibilities in shape. This is especially obvious when we compare the Type II frequency function with the polynomial of degree 4, in both of which the first through fourth moments were fitted.

Figure 4-1-3 illustrates two polynomials fitted by the method of moments to each of the two sets of observations. The figure belongs to the combination of the Two-Parameter Beta Method and the Curve Fitting Approach, Degree 3 Case, which will be introduced in the following chapter. 2,500 observations of θ , which were produced by the Monte Carlo method, were classified into two groups, i.e., the success and the failure groups for an unknown binary test item, item 4 . These two subsets of observations are shown in Figure 4-1-3 in the summarized form of frequency distributions, by thick and thin lines. For each subset, polynomials of degrees 3 and 4 were fitted by the method of moments, and are shown by a long, dashed line and a dotted line, respectively.

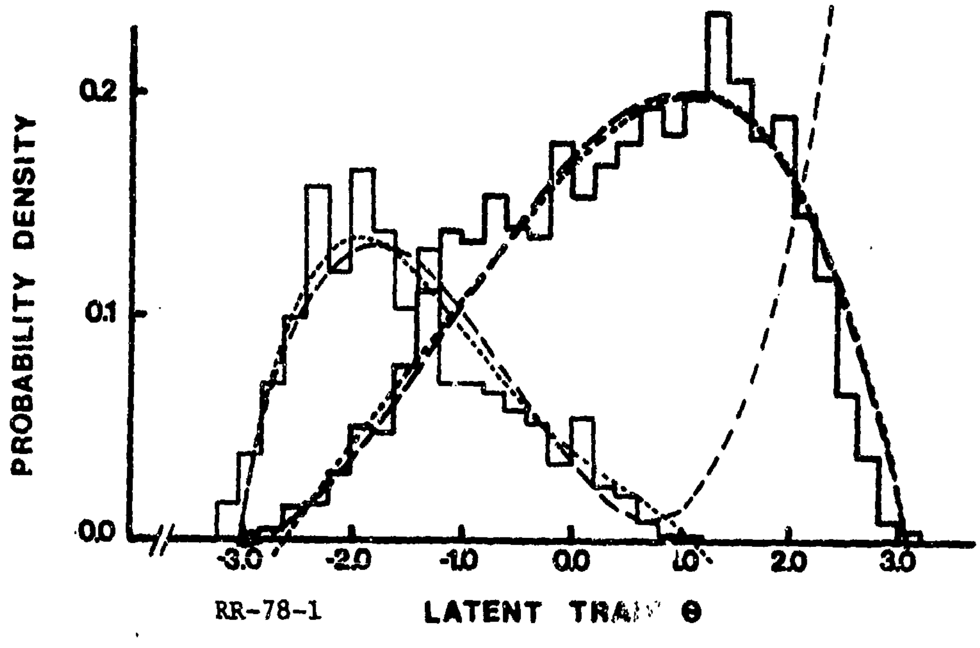


FIGURE 4-1-3

Relative Frequencies of θ Shared by the Success (Thick Line) and the Failure (Thine Line) Groups and the Corresponding Polynomials of Degree 3 (Long Dashes) and of Degree 4 (Dots) for Item 4 . Two-Parameter Beta Method and Curve Fitting Approach, Degree 3 Case.

(IV.2) Method of Moments As the Least Squares Solution for Fitting a Polynomial

Let $h(t)$ be any function of the variable t, which is defined in a closed interval, $[\underline{t}, \bar{t}]$, and is integrable in the Lebesgue sense and has the first m moments. This function $h(t)$ can be some specified mathematical function, or an empirically obtained function. Let α_i $(i=0,1,2,\ldots,m)$ be the i-th coefficient of the polynomial which can be written in the form

$$\sum_{i=0}^{m} \alpha_i t^i , \tag{4.4}$$

and is to be fitted to the function $h(t)$ following the least squares principle. We define Q such that

$$2Q = \int_{\underline{t}}^{\bar{t}} [h(t) - \sum_{i=0}^{m} \alpha_i t^i]^2 \, dt . \tag{4.5}$$

Differentiating Q with respect to α_r and setting the result equal to zero, we obtain

$$\frac{\partial Q}{\partial \alpha_r} = \int_{\underline{t}}^{\bar{t}} [h(t) - \sum_{i=0}^{m} \alpha_i t^i][-t^r] \, dt = 0 \tag{4.6}$$

and then

$$\int_{\underline{t}}^{\bar{t}} t^r h(t) \, dt = \int_{\underline{t}}^{\bar{t}} t^r \sum_{i=0}^{m} \alpha_i t^i \, dt , \tag{4.7}$$

for $r=1,2,\ldots,m$.

Thus it is obvious from (4.6) that the least squares principle requires the resultant polynomial of degree m to have the same 0-th through m-th moments as $h(t)$, which is nothing but the principle upon which the method of moments is based. From this result, it is obvious that both methods provide us with the same polynomial.

When the function $h(t)$ is observed only at N points of the variable t, as is often the case for an empirically observed function, we can replace (4.5) by

$$(4.8) \qquad 2Q = \sum_{k=1}^{N} \{[h(t_k) - \sum_{i=0}^{m} \alpha_i t_k^{i}] w(t_k)\}^2 ,$$

where $w(t_k)$ is some appropriately chosen weight for t_k. Differentiating (4.5) and setting the result equal to zero, we obtain

$$(4.9) \qquad \sum_{k=1}^{N} t_k^{r} h(t_k) w(t_k) = \sum_{k=1}^{N} t_k^{r} w(t_k) \sum_{i=0}^{m} \alpha_i t_k^{i} .$$

If the function $h(t)$ is continuous and we divide the interval $[\underline{t}, \bar{t}]$ into N subintervals, by the middle value theorem there exists at least one value, ζ_{kr}, in each subinterval $(\underline{t}_k, \bar{t}_k)$ which satisfies

$$(4.10) \qquad \int_{\underline{t}_k}^{\bar{t}_k} t^{r} h(t)\, dt = \zeta_{kr}^{r} h(\zeta_{kr})(\bar{t}_k - \underline{t}_k) ,$$

where

$$(4.11) \qquad \bar{t}_k = \underline{t}_{k+1}$$

for $k = 1, 2, \ldots, (N-1)$, and

$$(4.12) \qquad \begin{cases} \underline{t}_1 = \underline{t} \\ \bar{t}_N = \bar{t} \end{cases}$$

When the width of each subinterval is small enough, these $(m+1)$ values, ζ_{kr} $(r=0,1,2,\ldots,m)$, can be approximated by a single value, say the midpoint of the subinterval. Using such a value as t_k and the subinterval width as $w(t_k)$, we can approximate (4.2) by (4.6). If all the subinterval widths are equal, (4.6) is simplified to provide

$$(4.13) \qquad \sum_{k=1}^{N} t_k^{r} h(t_k) = \sum_{k=1}^{N} t_k^{r} \sum_{i=0}^{m} \alpha_i t_k^{i} .$$

(IV.3) Direct Use of the Least Squares Solution

We can rewrite (4.4) in the form

$$\mu'_s = \sum_{j=1}^{m+1} \alpha_{j-1} [j+s-1]^{-1} [t^{-j+s-1} - \underline{t}^{j+s-1}] \quad , \tag{4.14}$$

where $s=r+1=1,2,\ldots,m+1$, $j=i+1=1,2,\ldots,m+1$, and μ'_s is the (s-1)-th moment of t about the origin, defined by

$$\mu'_s = \int_{\underline{t}}^{\bar{t}} t^{s-1} h(t)\, dt \quad . \tag{4.15}$$

Let α be a column vector of order (m+1), whose j-th element is α_{j-1} , and μ' be a column vector of the same order whose s-th element is μ'_s . Thus we can rewrite (4.11) in the matrix notation to obtain

$$\mu' = A\alpha \quad , \tag{4.16}$$

where A is a symmetric matrix of order (m+1) whose sj-element is given by

$$[j+s-1]^{-1} [\bar{t}^{j+s-1} - \underline{t}^{j+s-1}] \quad . \tag{4.17}$$

The least squares solution for α is obtained, therefore, by

$$\hat{\alpha} = A^{-1}\mu' \quad . \tag{4.18}$$

For the purpose of illustration, the matrix A for $m = 2$ is shown below as an example.

$$A = \begin{bmatrix} (\bar{t} - \underline{t}) & (\bar{t}^2 - \underline{t}^2)/2 & (\bar{t}^3 - \underline{t}^3)/3 \\ (\bar{t}^2 - \underline{t}^2)/2 & (\bar{t}^3 - \underline{t}^3)/3 & (\bar{t}^4 - \underline{t}^4)/4 \\ (\bar{t}^3 - \underline{t}^3)/3 & (\bar{t}^4 - \underline{t}^4)/4 & (\bar{t}^5 - \underline{t}^5)/5 \end{bmatrix} \quad . \tag{4.19}$$

In practice, we usually use a greater value for m , and obtaining the inverse matrix of A will be the most intricate process of computation, and the availability of a package program for inversing a symmetric matrix wil) be of necessity.

(IV.4) <u>Solution by the Method of Moments</u>

Let R(t) be a half of the interval width for which the function h(t) is defined, and M(t) be the midpoint of the interval, such that

$$R(t) = (\bar{t} - \underline{t})/2 \tag{4.20}$$

and

$$M(t) = (\bar{t} + \underline{t})/2 \quad . \tag{4.21}$$

For convenience, we define a new variable t* by changing the origin of t to the midpoint of the interval $[\underline{t}, \bar{t}]$, i.e.,

$$t^* = t - M(t) \quad . \tag{4.22}$$

Thus the polynomial of degree m in t can be rewritten as a polynomial of the same degree in t* , or

$$\sum_{i=0}^{m} \alpha_i \, t^i = \sum_{i=0}^{m} a_i \, t^{*i} \quad , \tag{4.23}$$

with the relationship between the two sets of coefficients such that

$$\alpha_r \begin{cases} = a_r & \text{for } M(t) = 0 \\ = \sum_{i=r}^{m} (-1)^{i-r} a_i \binom{i}{r} [M(t)]^{i-r} \quad , & \text{otherwise,} \end{cases} \quad r=0,1,2,\ldots,m \ . \tag{4.24}$$

The following relationships hold between the moments about the midpoint M(t) and the coefficients a_r (r=1,2,...,m) .

$$\mu^*_{2g} = 2 \sum_{k=0}^{[m/2]} a_{2k} \, [2(g+k)+1]^{-1} \, [R(t)]^{2(g+k)+1} \qquad g=0,1,2,\ldots,[m/2] \ . \tag{4.25}$$

$$\mu^*_{2g+1} = 2 \sum_{k=0}^{[(m-1)/2]} a_{2k+1} \, [2(g+k+1)+1]^{-1} \, [R(t)]^{2(g+k+1)+1} \qquad g=0,1,2,\ldots,[(m-1)/2] \ . \tag{4.26}$$

In the above two equations, [ ] indicates the integer part of the number, and μ^*_{2g} and μ^*_{2g+1} indicate even and odd moments about the midpoint, M(t) , respectively.

Let $p=g+1$ and $q=k+1$. We define the following two symmetric matrices, $B_{(0)}$ and $B_{(1)}$, whose orders are both $(m+1)/2$ when m is odd, and $(m/2)+1$ and $(m/2)$ when m is even, respectively.

(4.27) $$B_{(0)} = \{ [R(t)]^{2(p+q)-3} [2(p+q)-3]^{-1} \} .$$

(4.28) $$B_{(1)} = \{ [R(t)]^{2(p+q)-1} [2(p+q)-1]^{-1} \} .$$

Let $\mu^*_{(0)}$ and $\mu^*_{(1)}$ be column vectors of the corresponding orders, such that

(4.29) $$\mu^*_{(0)} = \{ \mu^*_{2(p-1)} \}' , \qquad p=1,2,\ldots,[m/2]+1 ,$$

and

(4.30) $$\mu^*_{(1)} = \{ \mu^*_{2p-1} \}' , \qquad p=1,2,\ldots,[(m+1)/2] .$$

Let $a_{(0)}$ and $a_{(1)}$ denote the coefficient vectors of the corresponding orders, which can be written as

(4.31) $$a_{(0)} = \{ a_{2(q-1)} \}' , \qquad q=1,2,\ldots,[m/2]+1 ,$$

and

(4.32) $$a_{(1)} = \{ a_{2q-1} \}' , \qquad q=1,2,\ldots,[(m+1)/2] .$$

Thus we can rewrite (4.25) and (4.26) in the matrix notation such that

(4.33) $$\mu^*_{(0)} = 2B_{(0)}a_{(0)}$$

and

(4.34) $$\mu^*_{(1)} = 2B_{(1)}a_{(1)} \quad .$$

The coefficient matrices $a_{(0)}$ and $a_{(1)}$ are obtained, therefore, by

(4.35) $$a_{(0)} = (1/2)\ B^{-1}_{(0)}\ \mu^*_{(0)}$$

and

(4.36) $$a_{(1)} = (1/2)\ B^{-1}_{(1)}\ \mu^*_{(1)} \quad .$$

In practice, the computation is facilitated if we define two matrices, $C_{(0)}$ and $C_{(1)}$, of orders $[m/2]+1$ and $[(m+1)/2]$, respectively, such that

(4.37) $$C_{(0)} = \{\ [2(p+q)-3]^{-1}\ \}$$

and

(4.38) $$C_{(1)} = \{\ [2(p+q)-1]^{-1}\ \} \quad ,$$

which do not depend on a specific set of data but depend only upon the degree of the polynomial. From these two matrices, we can obtain the two matrices, $(1/2)\ C^{-1}_{(0)}$ and $(1/2)\ C^{-1}_{(1)}$. , and it is easily seen that $(1/2)\ B^{-1}_{(0)}$ and $(1/2)\ B^{-1}_{(1)}$ are obtained by dividing the element in the p-th row and q-th column of the corresponding matrices by $[R(t)]^{2(p+q)-3}$ and $[R(t)]^{2(p+q)-1}$, repectively, for every combination of p and q . The resultant sets of equations for obtaining the coefficients a_i are listed below for the polynomials of degrees 3, 4, 5, 6 and 7.

(i) Polynomial of Degree 3

$$
(4.39)\quad \begin{cases}
a_0 = [1.125\mu_0^*/R] - [1.875\mu_2^*/R^3] \\
a_1 = [9.375\mu_1^*/R^3] - [13.125\mu_3^*/R^5] \\
a_2 = [-1.875\mu_0^*/R^3] + [5.625\mu_2^*/R^5] \\
a_3 = [-13.125\mu_1^*/R^5] + [21.875\mu_3^*/R^7]
\end{cases}
$$

(ii) Polynomial of Degree 4

$$
(4.40)\quad \begin{cases}
a_0 = [1.7578125\mu_0^*/R] - [8.203125\mu_2^*/R^3] + [7.3828125\mu_4^*/R^5] \\
a_1 = [9.375\mu_1^*/R^3] - [13.125\mu_3^*/R^5] \\
a_2 = [-8.203125\mu_0^*/R^3] + [68.90625\mu_2^*/R^5] - [73.828125\mu_4^*/R^7] \\
a_3 = [-13.125\mu_1^*/R^5] + [21.875\mu_3^*/R^7] \\
a_4 = [7.3828125\mu_0^*/R^5] - [73.828125\mu_2^*/R^7] + [86.1328125\mu_4^*/R^9]
\end{cases}
$$

(iii) Polynomial of Degree 5

$$
(4.41)\quad \begin{cases}
a_0 = [1.7578125\mu_0^*/R] - [8.203125\mu_2^*/R^3] + [7.3828125\mu_4^*/R^5] \\
a_1 = [28.7109375\mu_1^*/R^3] - [103.359375\mu_3^*/R^5] + [81.2109375\mu_5^*/R^7] \\
a_2 = [-8.203125\mu_0^*/R^3] + [\ 68.90625\mu_2^*/R^5] - [73.828125\mu_4^*/R^7] \\
a_3 = [-103.359375\mu_1^*/R^5] + [442.96875\mu_3^*/R^7] - [378.984375\mu_5^*/R^9] \\
a_4 = [7.3828125\mu_0^*/R^5] - [73.828125\mu_2^*/R^7] + [86.1328125\mu_4^*/R^9] \\
a_5 = [81.2109375\mu_1^*/R^7] - [378.984375\mu_3^*/R^9] + [341.0859375\mu_5^*/R^{11}]
\end{cases}
$$

(iv) Polynomial of Degree 6

$$
(4.42)\quad \begin{cases}
a_0 \doteq [2.3925781\mu_0^*/R] - [21{,}5332031\mu_2^*/R^3] + [47.3730469\mu_4^*/R^5] \\
\qquad - [29.3261719\mu_6^*/R^7] \\
a_1 = [28.7109375\mu_1^*/R^3] - [103.359375\mu_3^*/R^5] + [81.2109375\mu_5^*/R^7] \\
a_2 \doteq [-21{,}5332031\mu_0^*/R^3] + [348{,}8378906\ {}_2^*/R^5] \\
\qquad - [913{,}6230469\mu_4^*/R^7] + [615{,}8496094\mu_6^*/R^9] \\
a_3 = [-103.359375\mu_1^*/R^5] + [442.96875\mu_3^*/R^7] - [378.984375\mu_5^*/R^9] \\
a_4 \doteq [47.3730469\mu_0^*/R^5] - [913{,}6230469\mu_2^*/R^7] \\
\qquad + [2605{,}5175781\mu_4^*/R^9] - [1847{,}5488281\mu_6^*/R^{11}] \\
a_5 = [81.2109375\mu_1^*/R^7] - [378.984375\mu_3^*/R^9] + [341.0859375\mu_5^*/R^{11}] \\
a_6 \doteq [-29.3261719\mu_0^*/R^7] + [615.8496094\mu_2^*/R^9] \\
\qquad - [1847.5488281\mu_4^*/R^{11}] + [1354.8691406\mu_6^*/R^{13}]
\end{cases}
$$

(v) Polynomial of Degree 7

$$
(4.43)\quad
\left\{
\begin{aligned}
a_0 &\doteq [2.3925781\mu_0^*/R] - [21.5332031\mu_2^*/R^3] + [47.3730469\mu_4^*/R^5] \\
&\qquad - [29.3261719\mu_6^*/R^7] \\
a_1 &\doteq [64.5996094\mu_1^*/R^3] - [426.3574219\mu_3^*/R^5] \\
&\qquad + [791.8066406\mu_5^*/R^7] - [439.8925781\mu_7^*/R^9] \\
a_2 &\doteq [-21.5332031\mu_0^*/R^3] + [348.8378906\ {}_2^*/R^5] \\
&\qquad - [913,6230469\mu_4^*/R^7] + [615,8496094\mu_6^*/R^9] \\
a_3 &\doteq [-426.3574219\mu_1^*/R^5] + [3349.9511719\mu_3^*/R^7] \\
&\qquad - [6774.3457031\mu_5^*/R^9] + [3959.0332031\mu_7^*/R^{11}] \\
a_4 &\doteq [47.3730469\mu_0^*/R^5] - [913.6230469\mu_2^*/R^7] \\
&\qquad + [2605.5175781\mu_4^*/R^9] - [1847.5488281\mu_6^*/R^{11}] \\
a_5 &\doteq [791.8066406\mu_1^*/R^7] - [6774.3457031\mu_3^*/R^9] \\
&\qquad + [14410.8808594\mu_5^*/R^{11}] - [8709.8730469\mu_7^*/R^{13}] \\
a_6 &\doteq [-29.3261719\mu_0^*/R^7] + [615.8496094\mu_2^*/R^9] \\
&\qquad - [1847.5488281\mu_4^*/R^{11}] + [1354.8691406\mu_6^*/R^{13}] \\
a_7 &\doteq [-439.8925781\mu_1^*/R^9] + [3959.0332031\mu_3^*/R^{11}] \\
&\qquad - [8709.8730469\mu_5^*/R^{13}] + [5391.8261719\mu_7^*/R^{15}]
\end{aligned}
\right.
$$

(For simplicity, in the above equations, R is used instead of $R(t)$.)

From the values of a_i's thus obtained and the midpoint $M(t)$, we can find out the values of the coefficients, α_i's, by means of (4.24).

(IV.5) Expanded Use of the Method of Moments

As we have observed in the preceding sections, the method of moments for fitting a polynomial can be considered another procedure for the least squares solution. It has a definite advantage over the direct least squares solution, since the computation of the coefficients, α_i $(i=0,1,2,\dots,m)$, can be done by the application of straight-forward algebra, while the direct procedure for the least squares solution involves the inversion of the matrix A.

This fact implies that we can adopt the method of moments for fitting a polynomial for the approximation to any target function, which is not necessarily a density function or a frequency distribution. In fact, in the present study, we used the method for approximating the square root of the test information function of the

Old Test, among others, which facilitated the transformation of ability θ to τ. The rationale behind this method will be described in the following chapter.

(IV.6) Selection of the Interval

When we fit a polynomial to a frequency distribution or a set of observations, the selection of the interval is more or less automatical. When we use the method of moments for fitting a polynomial to a function other than those, however, the goodness of fit of the polynomial to the target function depends largely upon our selection of the interval.

Figure 4-6-1 illustrates such a situation. In this figure, the square root of the test information function, $[I(\theta)]^{1/2}$, of Subtest 1 is drawn by a solid line. The other two dashed and dotted curves are the polynomials of degree 7 obtained by the method of moments, using the intervals of θ, [-3.0, 3.0] and

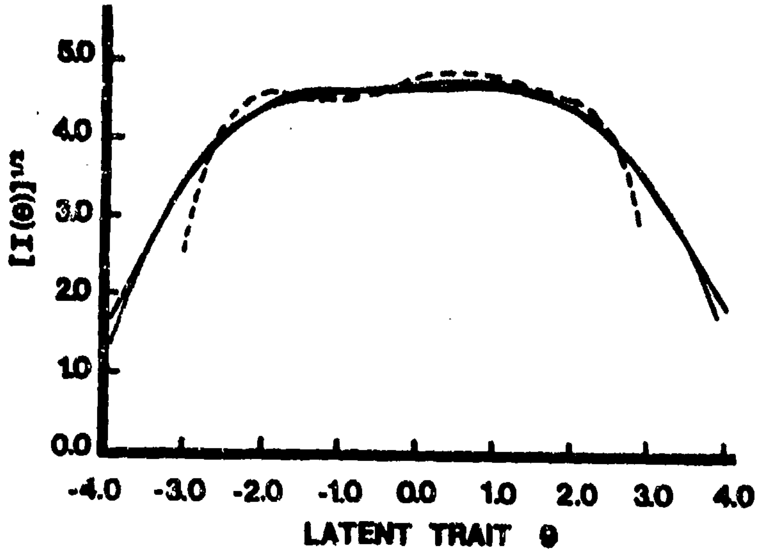


FIGURE 4-6-1

Square Root of the Test Information Function of Subtest 1, $[I(\theta)]^{1/2}$, (Solid Line) and the Polynomials of Degree 7, Which were Fitted by the Method of Moments with [-3.0, 3.0] (Dashes) and [-4.0, 4.0] (Dots) as the Interval of θ, Respectively.

[-4.0, 4.0] , respectively. We can see that the latter polynomial fits much better than the former to the target function. This implies that, although the interval of ability θ of our interest is even a little smaller than [-3.0, 3.0] , in order to obtain a polynomial which fits to the target function in this interval, we must use a larger interval such as [-4.0, 4.0] .

We cannot generalize this result too much, however. Figure 4-6-2 presents a similar set of curves for Subtest 2 . It is noted that, while the fit is better for the polynomial obtained by using the interval, [-4.0, 4.0] , than the one obtained by using the interval, [-3.0, 3,0] , in the former situation there still is a substantial discrepancy form the target function.

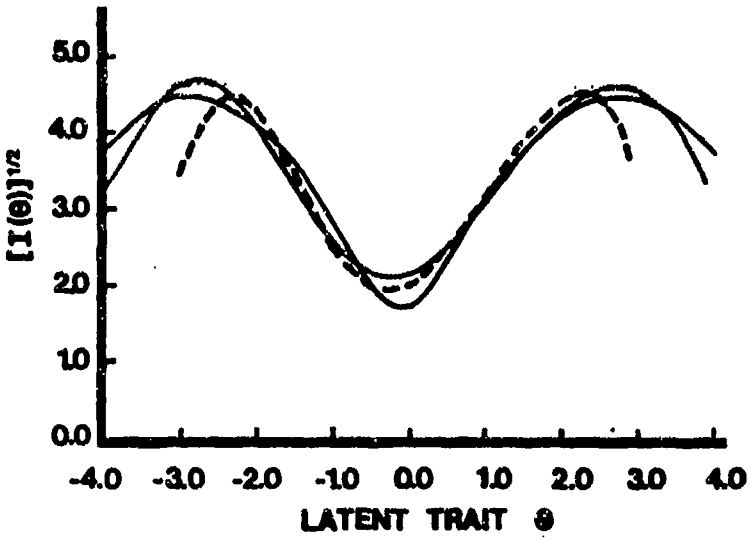


FIGURE 4-6-2

Square Root of the Test Information Function of Subtest 2 , $[I(\theta)]^{1/2}$, (Solid Line) and the Polynomials of Degree 7 , Which were Fitted by the Method of Moments with [-3.0, 3.0] (Dashes) and [-4.0, 4.0] (Dots) as the Interval of θ , Respectively.

Figure 4-6-3 presents the result obtained by using the three subintervals of [-4.0, 4.0] , with θ = -1.5 and 0.5 as the cutting points. These three polynomials are uniformly of degree

4 . We can see that, together, they fit very well to the target function. This is another way of using the method of moments.

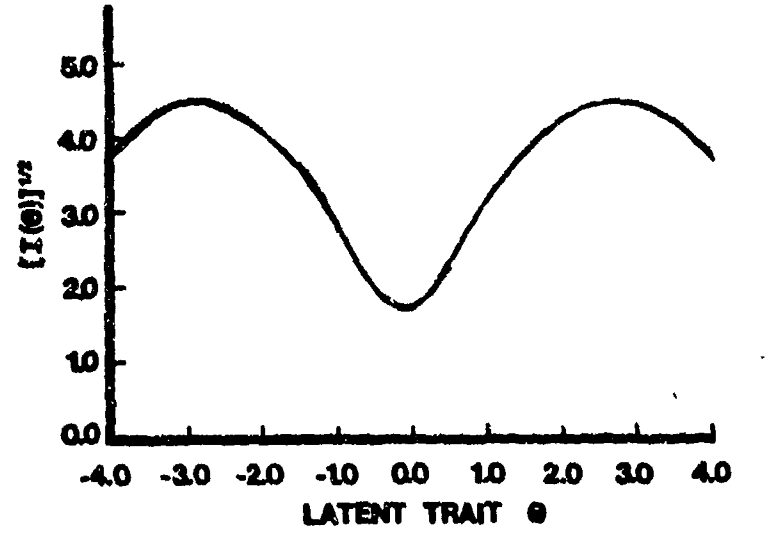


FIGURE 4-6-3

Square Root of the Test Information Function of Subtest 2, $[I(\theta)]^{1/2}$, (Solid Line) and the Three Polynomials of Degree 4 (Dots), Which Were Fitted by the Method of Moments Using the Three Subintervals of θ .

The use of subintervals may be effective when we apply the method of moments for fitting polynomials to relatively smooth mathematical functions. The same is not necessarily true, however, if we use the method of moments for empirical data. Figure 4-6-4 illustrates such examples. Our data are again the set of five hundred maximum likelihood estimates $\hat{\theta}_s$, and, in the first graph, it was reclassified into the lower and upper subsets of 250 observations each, and, in the second graph, in a similar manner, it was divided into five subsets of 100 observations each. The polynomials shown in these two graphs are uniformly of degree 4 . We can see that neither result is appropriate for us to use as the estimated density function, $\hat{g}(\hat{\theta})$.

To conclude, the selection of the interval or intervals is very important in order to use the method of moments for fitting a polynomial or polynomials successfully, and we must make a good judgment in each situation considering the expected shape of the target function, and the nature of our data.

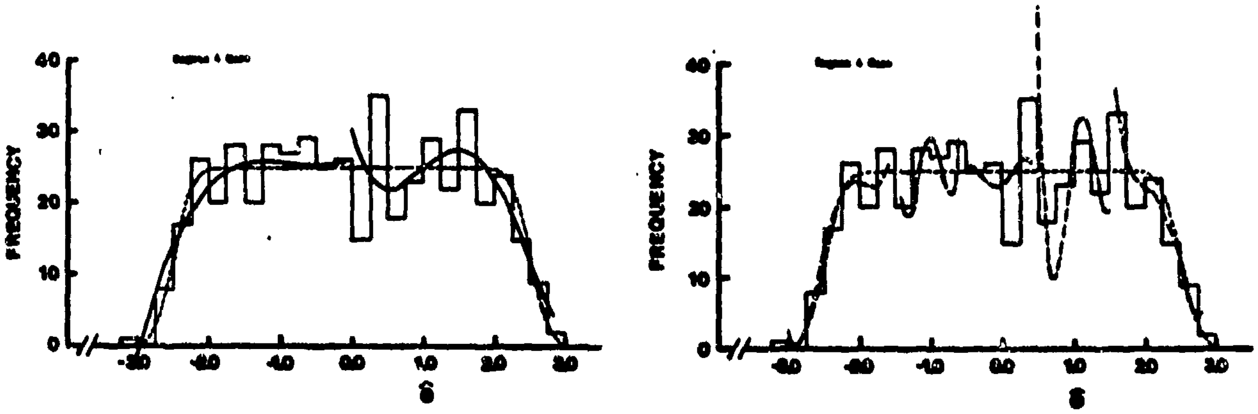


FIGURE 4-6-4

Polynomial Approximations of the Density Function $g(\theta)$ of the Set of Five Hundred Maximum Likelihood Estimates $\hat{\theta}_s$ Obtained upon the Original Old Test by the Method of Moments, by Dividing the Total Set into Two Subsets (Left) and into Five Subsets (Right).

There are many examples other than those illustrated here, and it is recommended that the reader refers to the research report, RR-79-2, and many others such as RR-78-1, RR-80-2 and RR-80-4.

(IV.7) Comparison of the Results Obtained by the Method of Moments and by the Direct Least Squares Procedure

Comparison of the polynomials obtained by the method of moments and by the direct least squares method was made by using the standard normal distribution function as the target function (cf. RR-79-2). It was made by changing the interval of θ for which these methods are applied, and, as is expected, in most cases the resultant two polynomials are identical.

There are somewhat different results, however. Figure 4-7-1 presents such an example. In this figure, the resultant polynomial obtained by the method of moments is plotted by dots, and the one obtained by the direct least squares method is shown by short dashes. In both cases the interval of θ , [-6.0, 6.0] , was adopted. It

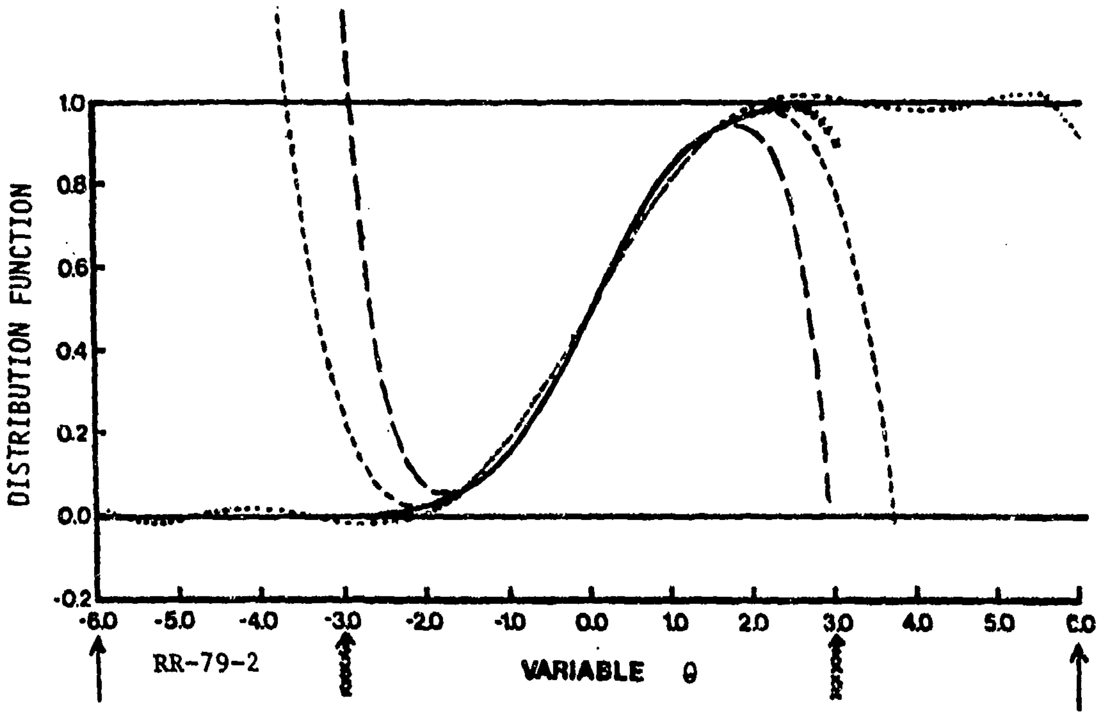


FIGURE 4-7-1

Polynomials of Degree 7 Obtained by the Method of Moments (Dots) and the Least Squares Solution (Short Dashes) , with the Interval, [-6.0, 6.0] , and the Taylor's Series (Long Dashes), Approximating the Standard Normal Distribution Function (Solid Line). Those Obtained by the First Two Methods Using the Interval, [-3.0, 3.0] , Are Also Plotted (Crosses).

is noted that, while the result obtained by the method of moments fits to the target function reasonable well for the total interval of θ , the one obtained by the direct least squares solution diverts, quickly, from the target function outside the interval, (-2.0, 2.0) . This diversion comes from the limitation of the capacity of the computer in inverting the matrix A . This example also suggests, therefore, that it is wise for us to use the method of moments instead of the direct least squares method. In the same figure, the corresponding two polynomials obtained by using the interval of θ , [-3.0, 3.0] , are also plotted. Since they are identical, they are drawn together by crosses, and only for the interval where the curves divert from the target function.

We recall that there is another type of polynomials which are obtained by Taylor's series. Using Hermite polynomials (Kendall and Stuart, 1963), we can write for the Taylor's series for the standard normal distribution function such that

$$(4.44) \qquad N(0,1) \doteq 0.500000 + 0.398942\theta - 0.0664903\theta^{3} + 0.00997355\theta^{5} \\ - 0.00118732\theta^{7} + 0.000115434\theta^{9} - \\ - 0.00000944465\theta^{11} + \ldots$$

The resultant polynomial of degree 7 is drawn by longer dashes in Figure 4-7-1. It is noted that the fit of this polynomial to the target function is better for the interval of θ , (-1.7, 1.7) , but outside of this interval it diverts from the target function quickly. This is a common tendency over the results of different degrees of polynomials (cf. RR-79-2).

References

[1] Elderton, W. P. and N. L. Johnson. Systems of frequency curves. Cambridge University Press, 1969.

[2] Kendall, M. G. & Stuart, A. The advanced theory of statistics (Vol. 1). New York: Hafner, 1963.

V Estimation of the Operating Characteristics of the Discrete Item Responses and That of Ability Distributions: II

In the present chapter, following Chapter 3, we shall continue integrating the rationale and findings of this part of the research. Throughout this process, the method of moments will frequently be used, especially, for fitting polynomials. The reasons for the choice of the polynomial in preference to the other functions were described in the preceding chapter. Among others, it provides us with the least squares solution.

(V.1) Estimated Operating Characteristics Which Are Directly Observable from Our Calibration Data

Since our data are simulated data, the proportion correct for each of the ten unknown, binary items (cf. Section III.6) is directly observable. Figure 5-1-1 illustrates two sets of the proportion correct for item 6, by solid and dashed lines, respectively, together with the theoretical item characteristic function. The subinterval widths used for these two curves are 0.05

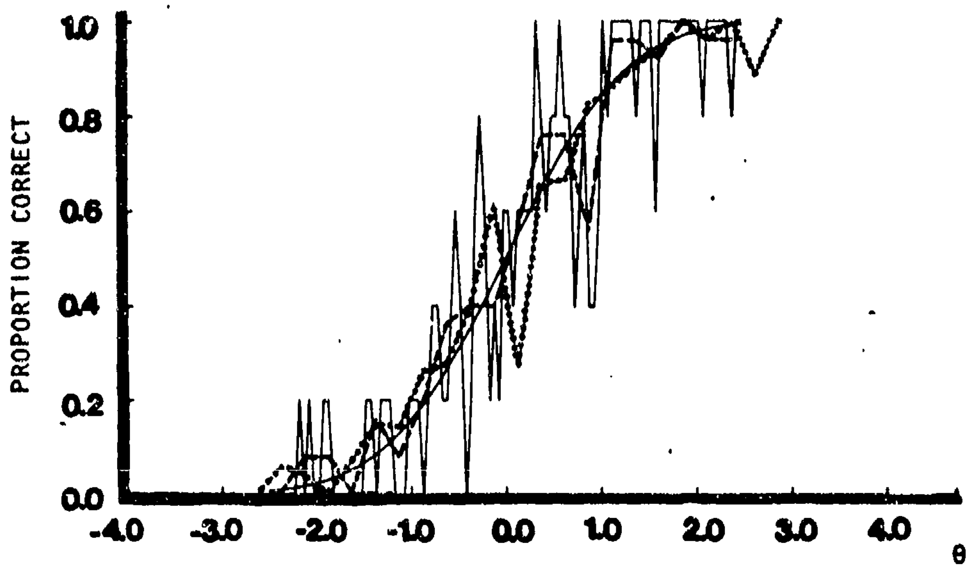


FIGURE 5-1-1

Proportion Correct for Item 6 Using the Subinterval Width 0.05 (Solid Line) and 0.25 (Dashed Line), and the Similar Result Obtained by Using the Maximum Likelihood Estimate $\hat{\theta}$ Instead of Ability θ and the Subinterval Width 0.25 (Dotted Line), Together with the Item Characteristic Function (Solid Curve).

and 0.25 , respectively. Thus, in the first case, five hypothetical examinees sharing the same position (cf. Section III.3) makes the total frequency for each subinterval of θ , and, in the second case, twenty-five examinees sharing five adjacent positions makes the total frequency. We can see from these results that they are by no means good approximations to the theoretical item characteristic function, because of their large fluctuations. The reason is that we have only five hundred hypothetical examinees in our calibration data.

It should be noted that these two curves in Figure 5-1-1 are not observable, if our calibration data are empirical data. In practice, the closest we can get from our empirical data is, therefore, the proportion correct based upon the maximum likelihood estimate $\hat{\theta}$, instead of ability θ itself. This third proportion correct for item 6 is also plotted in Figure 5-1-1 by a dotted line, using the set of five hundred maximum likelihood estimates obtained upon our original Old Test (cf. Section III.3). The subinterval width for this proportion correct is 0.25 , as was the case with the second curve based upon ability θ . Again, we can see that the fluctuations from the true item characteristic function are large.

As was poin:ed out in Section III.7, the use of indirect information obtainable from our calibration data will ameliorate the situation. If our results provide us with better approximations to the theoretical item characteristic function than those three curves do, therefore, we shall content ourselves by deciding that our methods are successful.

(V.2) Necessary Correction for the Scale of the Maximum Likelihood Estimate When Used As a Substitute for Ability Scale

It is commonly taken for granted that, whenever the scale of the maximum likelihood estimate is available, it can directly be used as the substitute for the ability scale. The reader may wonder, therefore, why we need an elaborated process of estimating the operating characteristics when the set of maximum likelihood estimates of ability is available. If our calibration data contain only several hundred examinees, because of the sampling fluctuations they cannot provide us with a good approximation to the theoretical operating characteristics, as we have seen in the example given in the preceding section.

Our next question will be: Is it justifiable to use the scale of maximum likelihood estimate and its proportion correct for the estimated item characteristic function when we have a large set of calibration data, like those based upon twenty thousand examinees? The answer still must be "No," or "Not without some modification."

Let us assume that our Old Test provides us with the approximate unbiasedness of the maximum likelihood estimate θ , and the normality for its conditional distribution, given ability θ , for the interval of θ , $(\underline{\theta},\bar{\theta})$, of our interest. Thus we can write

$$(5.1)\qquad E(\hat{\theta}|\theta) = 0 \ .$$

From (5.1), we obtain for the expectation of $\hat{\theta}$ such that

$$(5.2)\qquad E(\hat{\theta}) = \int_{\underline{\theta}}^{\bar{\theta}} E(\hat{\theta}|\theta)\ f(\theta)\ d\theta = E(\theta) \ .$$

By virtue of the binomial law, we have, from (5.2), for the m-th moment of $\hat{\theta}$ about the mean

$$(5.3)\qquad E[\hat{\theta}-E(\hat{\theta})]^{m} = \sum_{r=0}^{m} \binom{m}{r} E[\{\theta-E(\theta)\}^{m-r}\ E\{(\hat{\theta}-\theta)^{r}|\theta\}] \ .$$

From (5.3), we can write for the specific cases where $m = 2$, 3 and 4 ,

$$(5.4)\qquad \mathrm{Var.}(\hat{\theta}) = \mathrm{Var.}(\theta) + E[\mathrm{Var.}(\hat{\theta}|\theta)] \ ,$$

$$(5.5)\qquad E[\{\hat{\theta}-E(\hat{\theta})\}^{3}] = E[\{\theta-E(\theta)\}^{3}] + 3E[\{\theta-E(\theta)\}\ \mathrm{Var.}(\hat{\theta}|\theta)]$$

and

$$(5.6) \qquad E[\{\hat{\theta}-E(\hat{\theta})\}^4] = E[\{\theta-E(\theta)\}^4] + 6E[\{\theta-E(\theta)\}^2\{E\{\mathrm{Var.}(\hat{\theta}|\theta)\}] + E[(\hat{\theta}-\theta)^4|\theta] .$$

The above results imply that the distribution of the maximum likelihood estimate $\hat{\theta}$ is different from that of ability θ , and, above all, it has a larger variance. Since the proportion correct is the ratio of two such distributions, these results indicate that it contains a bias in itself.

The correction for this distortion can be made in the following way. Let us assume, tentatively, that the square root of the test information function of our Old Test is approximately constant for the interval, $(\underline{\theta},\bar{\theta})$, as is the case with our original Old Test (cf. Section III.4). Then the conditional distribution of $\hat{\theta}$, given θ , is approximately $N(\theta,\sigma)$, where σ is the reciprocal of the constant square root of the test information function, $[I(\theta)]^{-1/2}$. Under this condition, the formulas (5.4) through (5.6) can be simplified to provide us with

$$(5.7) \qquad \mathrm{Var.}(\hat{\theta}) = \mathrm{Var.}(\theta) + \sigma^2 ,$$

$$(5.8) \qquad E[\{\hat{\theta}-E(\hat{\theta})\}^3] = E[\{\theta-E(\theta)\}^3]$$

and

$$(5.9) \qquad E[\{\hat{\theta}-E(\hat{\theta})\}^4] = E[\{\theta-E(\theta)\}^4] + 6\sigma^2\,\mathrm{Var.}(\theta) + 3\sigma^4 .$$

Thus the distribution of $\hat{\theta}$ has the same mean and the third moment about mean as that of θ .

The regression of ability θ on the maximum likelihood estimate $\hat{\theta}$ is given in Chapter 3 as (3.22) . To reproduce it, we have

$$(5.10) \qquad E(\theta|\hat{\theta}) = \hat{\theta} + \sigma^2 \frac{d}{d\hat{\theta}} \log g(\hat{\theta}) ,$$

where

$$(5.11)\qquad g(\theta) = \int_{\underline{\theta}}^{\bar{\theta}} \psi(\hat{\theta}|\theta)\ f(\theta)\ d\theta \ .$$

Note that the regression, which is given by (5.10), is not necessarily linear, although that of the maximum likelihood $\hat{\theta}$ on ability θ is. (5.10) can be evaluated if we approximate the density function, $g(\hat{\theta})$, by, say, a polynomial obtained by the method of moments. We can shift the value of $\hat{\theta}$, therefore, to $E(\theta|\hat{\theta})$, so that we make the proportion correct for a specific value of $\hat{\theta}$ the function of the corresponding value of $E(\theta|\hat{\theta})$.

When the square root of the test information function of our Old Test is not constant, as is the case with each of the nine subtests of our original Old Test, we cannot directly apply the above method. In such a case, we must transform θ to τ, follow the whole process by using τ instead of θ, and then retransform τ to θ. The rationale behind this transformation is given in Section III.8 , and its actual procedure, using the approximation to the square root of the test information function by a polynomial obtained by the method of moments, will be given in the following section.

The observations made in this section have nothing to do with our methods and approaches for estimating ability distributions and the operating characteristics of discrete item responses, however. In the present study, either the conditional distribution of ability θ, given its maximum likelihood estimate $\hat{\theta}$, or the bivariate distribution of θ and $\hat{\theta}$ is approximated from our calibration data. This does not include, therefore, the direct frequency ratios of the maximum likelihood estimate, $\hat{\theta}$.

(V.3) Transformation of θ to τ Using the Method of Moments for Fitting a Polynomial

The rationale behind the transformation of θ to τ is given in Section III.8 . This process will be simplified if we use the approximation to the square root of the test information function of our Old Test by a polynomial fitted by the method of moments. In so doing, the right selection of the interval of θ for which the method of moments is applied is very important, as was explained in Section IV.6..

We can write

$$(5.12) \qquad [I(\theta)]^{1/2} \doteq \sum_{k=0}^{m} \alpha_k \theta^k ,$$

where m is the degree of the polynomial we wish to obtain. Substituting (5.12) into (3.40), we obtain

$$(5.13) \qquad \tau \doteq C^{-1} \sum_{k=0}^{m} \alpha_k (k+1)^{-1} \theta^{k+1} + d$$

$$= \sum_{k=0}^{m+1} \alpha_k^* \theta^k ,$$

where

$$(5.14) \qquad \alpha_k^* \begin{cases} = d & k = 0 \\ = (Ck)^{-1} \alpha_{k-1} & k = 1,2,\ldots,m+1 \end{cases} .$$

Thus the transformation of θ to τ can be made through another polynomial of degree $(m+1)$. Considering that (3.40) includes a tedious numerical process of integrating $[I(\theta)]^{1/2}$, the straight forward method given by (5.13) and (5.14) will save us a substantial amount of time and labor.

Figure 5-3-1 presents the transformation of θ to τ obtained by this method, for Subtests 1 and 2, to represent those for the nine subtests. In all of these nine cases, the interval, $[-4.0, 4.0]$, was used in applying the method of moments.

Figure 5-3-2 presents the resultant square root of the test information function, $[I^*(\tau)]^{1/2}$, for Subtests 1 and 2. As is

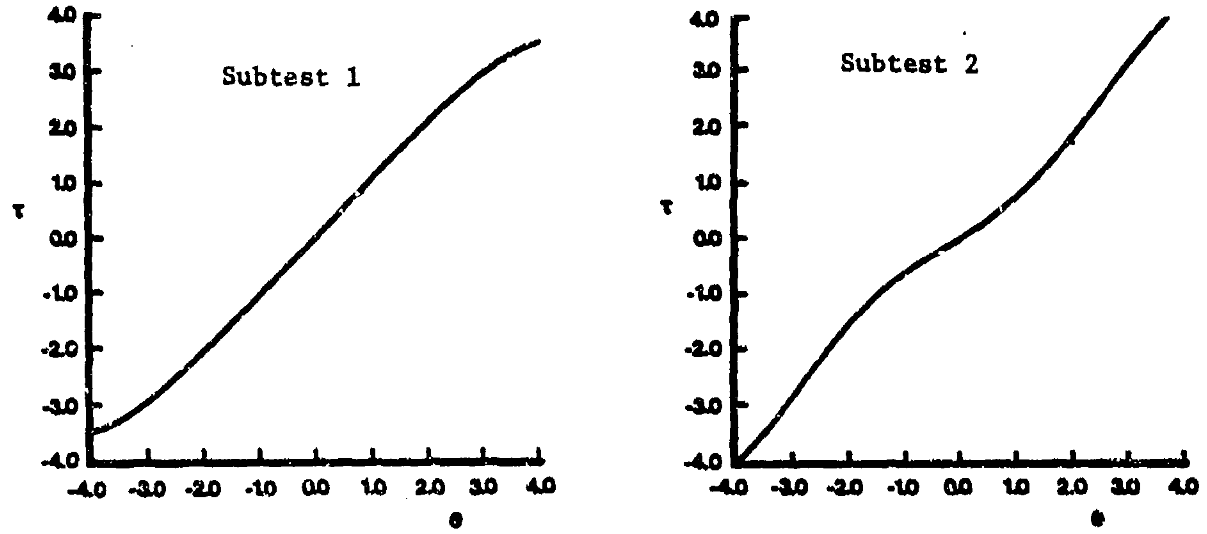


FIGURE 5-3-1

Transformation of θ to τ for Subtests 1 and 2 .

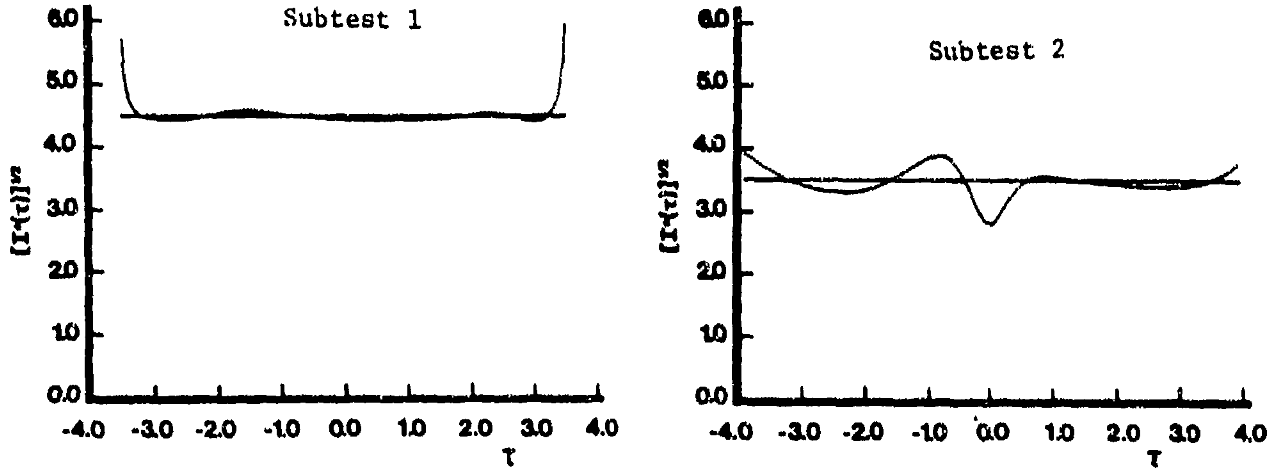


FIGURE 5-3-2

Square Root of Test Information Function, $[I^*(\tau)]^{1/2}$, and the Target Constant C for Subtests 1 and 2 .

expected from the different degrees of fitness of the polynomials to the respective $[I(\theta)]^{1/2}$'s in these two cases, which are shown in Figures 4-6-1 and 4-6-2 of Section IV.6, respectively, the resultant $[I^*(\tau)]^{1/2}$ for Subtest 1 is closer to the target constant than the one for Subtest 2. For all the other seven subtests, the result is similar to either one of these two results, or their fitness is somewhere between the two.

(V.4) Classification of Methods and Approaches

Various methods and approaches for estimating the operating characteristics of discrete item responses, and for estimating ability distributions, were developed in the present study. For convenience, by a method we mean a way of approximating the conditional density function of ability θ or τ, given its maximum likelihood estimate $\hat{\theta}$ or $\hat{\tau}$, and by an approach we mean a way of producing the ability distributions of separate discrete response groups, and hence the operating characteristics (cf. Section III.1). They are summarized as follows.

(A) Methods

(i) Pearson System Method
(ii) Two-Parameter Beta Method
(iii) Normal Approach Method

(B) Approaches

(i) Bivariate P.D.F. Approach
(ii) Histogram Ratio Approach
(iii) Curve Fitting Approach
(iv) Conditional P.D.F. Approach
 (a) Simple Sum Procedure
 (b) Weighted Sum Procedure
 (c) Proportioned Sum Procedure

Prior to the present study, the author had developed a method (Samejima, 1977) of estimating the operating characteristics of discrete item responses, which, later, was called Normal Approximation Method. With the classification given above, this method belongs to the Bivariate P.D.F. Approach. Although it had been developed before the author started the present study, a brief description of this approach will be given in Section V.5, so that the reader will understand its characteristics and the differences from the combinations of a method and an approach, which are the main products of the present study.

(V.5) Normal Approximation Method

Let $h(\theta)$ be a linear function of $\hat{\theta}$, which minimizes the quantity Q, such that

$$Q = E[\theta - h(\hat{\theta})]^2 \quad . \tag{5.15}$$

We obtain

$$h(\hat{\theta}) = \text{Cov.}(\theta,\hat{\theta})\ [\text{Var.}(\hat{\theta})]^{-1}\ [\hat{\theta} - E(\hat{\theta})] + E(\theta) \quad , \tag{5.16}$$

where $\text{Cov.}(\theta,\hat{\theta})$ denotes the covariance of ability θ and its maximum likelihood estimate $\hat{\theta}$.

When the square root of the test information function of our Old Test is approximately constant for the interval of θ of our interest, as is the case with our original Old Test, we can write from (5.7)

$$\text{Cov.}(\theta,\hat{\theta}) = \text{Var.}(\theta) = \text{Var.}(\hat{\theta}) - \sigma^2 \quad . \tag{5.17}$$

Substituting (5.17) into (5.16) and rearranging, we obtain

$$\begin{aligned} h(\hat{\theta}) &= [1-\sigma^2\{\text{Var.}(\hat{\theta})\}^{-1}][\hat{\theta} - E(\hat{\theta})] + E(\theta) \\ &= [1-\sigma^2\ \text{Var.}(\hat{\theta})\}^{-1}]\hat{\theta} + \sigma^2[\text{Var.}(\hat{\theta})]^{-1}\ E(\hat{\theta}) \\ &= \beta\hat{\theta} + \alpha \quad . \end{aligned} \tag{5.18}$$

From this result, it is obvious that the two coefficients, α and β, can be estimated from the set of maximum likelihood estimates. When the joint distribution of θ and $\hat{\theta}$ is normal, this function, $h(\hat{\theta})$, becomes the regression of θ on $\hat{\theta}$. In such a case, the conditional distribution of θ, given $\hat{\theta}$, is normal, with the common conditional variance such that

$$(5.19)\qquad \mathrm{Var.}(\theta|\hat{\theta}) = \sigma^2[1-\sigma^2\{\mathrm{Var.}(\hat{\theta})\}^{-1}] \ .$$

In the Normal Approximation Method, a bivariate normal distribution is assumed for the joint distribution of θ and $\hat{\theta}$ for each subpopulation of examinees who share the same discrete item response to an unknown test item. With our calibration data, there are two groups of examinees, i.e., the success and failure groups, for each of the ten binary test items (cf. Section III.6).

For each of the five hundred maximum likelihood estimates, $\hat{\theta}_s$, using the Monte Carlo method, a single value of θ was calibrated. Let $\tilde{\theta}$ denote this calibrated value of θ . Then we have two subtests of $\tilde{\theta}$, for the success and failure groups of item h , respectively. The ratio of the frequency distribution of the success group to the sum of the two frequency distributions makes the estimated item characteristic function of item h .

Figure 5-5-1 presents by hollow circles the estimated item characteristic function of item 6, thus obtained by using 0.25 as the subinterval width of frequency distributions of θ . In the same figure, also presented by solid triangles and hollow squares are the estimated item characteristic functions obtained by producing five and ten $\tilde{\theta}$'s for each of the five hundred maximum likelihood estimates, $\hat{\theta}_s$, respectively, in order to increase the accuracy of estimation. We can see that even with the five hundred $\tilde{\theta}$'s , the estimated item characteristic function is fairly close to the theoretical item characteristic function, and it becomes closer when we increase the number of $\tilde{\theta}$'s to 2,500 and to 5,000.

When our Old Test does not have a constant square root of the test information function for the interval of θ of our interest, as is the case with the nine subtests of the original Old Test, we can transform θ to τ and follow the same process. To obtain the estimated operating characteristics, we can

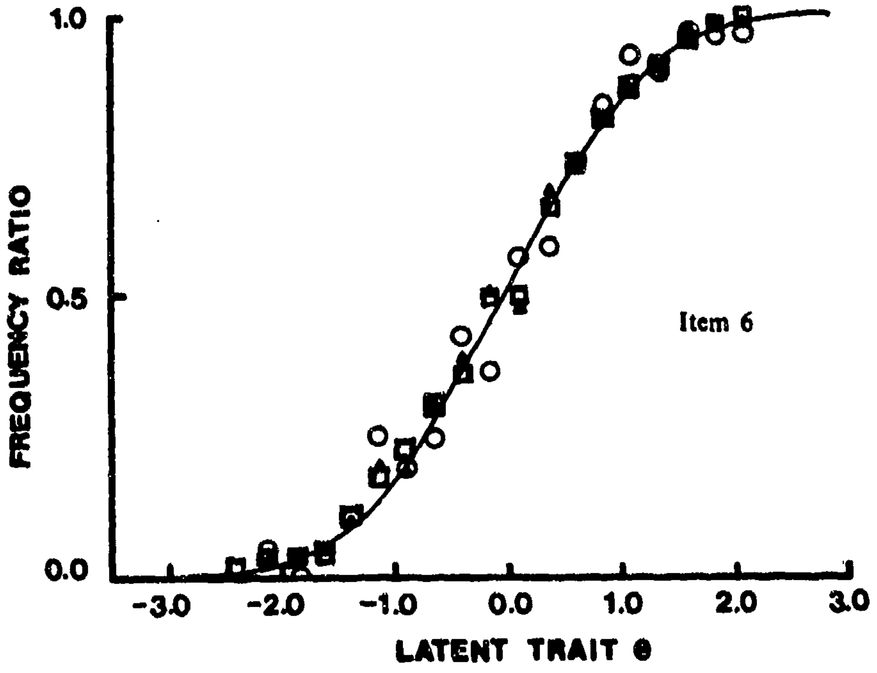


FIGURE 5-5-1

Estimated Item Characteristic Functions of Item 6 Based upon 500 $\tilde{\theta}$'s (Hollow Circles), upon 2,500 $\tilde{\theta}$'s (Solid Triangles) and upon 5,000 $\tilde{\theta}$'s (Hollow Squares), by the Normal Approximation Method, Using the Original Old Test.

retransform τ to θ after the process has been completed (cf. Sections III.8 and V.3) .

(V.6) Approximation to the Density Function of the Maximum Likelihood Estimate by a Polynomial Obtained by the Method of Moments

It is noted that, in the Normal Approximation Method, the marginal density function, $g(\hat{\theta})$, is totally unused. In contrast to this fact, in the present study, we make the full use of this marginal density function. In so doing, we approximate $g(\hat{\theta})$ or $g(\hat{\tau})$, depending upon the necessity of the transformation θ to τ , by a polynomial obtained by the method of moments. An example of this approximation was already given in Section IV.1, as Figure

4-2-1 . In this example, three different polynomials of degrees 3, 4, and 5 were fitted to the total set of five hundred maximum likelihood estimates $\hat{\theta}_s$, which are based upon the original Old Test. These three different situations are called Degree 3 Case, Degree 4 Case and Degree 5 Case, respectively.

Figure 5-6-1 presents another example of approximating the density function by a polynomial obtained by the method of moments. In this example, however, the target density function is divided into two portions, which belong to those who answered item h correctly and those who did not, respectively, and three polynomials of degrees 3, 4 and 5 were fitted to each portion. The result illustrated here is for item 6, and the original Old Test was used for producing the five hundred maximum likelihood estimates.

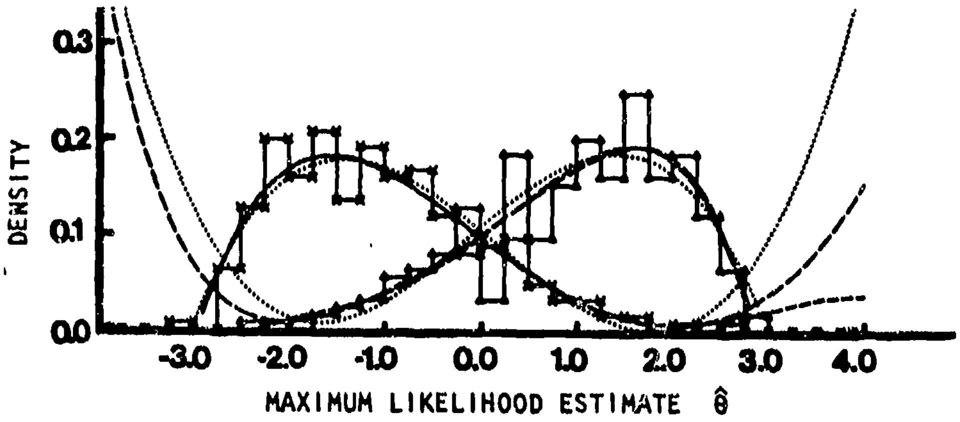


FIGURE 5-6-1

Approximations to the Two Portions of the Density Function, $g(\hat{\theta})$, for the Success and Failure Groups of Item 6, Respectively, by Polynomials of Degree 3 (Dots), of Degree 4 (Short Dashes) and of Degree 5 (Long Dashes) Obtained by the Method of Moments. Maximum Likelihood Estimates Are Based upon the Original Old Test, and Are Shown As Two Histograms.

To distinguish the two subset. of the maximum likelihood estimates from each other, the histogram of $\hat{\theta}_s$ for the failure group is marked with crosses, and the one for the success group is marked with solid triangles. The two polynomials of degree 3 are drawn by dotted lines, those of degree 4 are plotted by short dashed lines, and those of degree 5 are drawn by long dashed lines. This is an example chosen from those which are used in the Bivariate P.D.F. Approach, which will be introduced in Section V.10 . From these approximated density functions, we can obtain the estimated conditional moments of ability θ , given its maximum likelihood estimate $\hat{\theta}$, through the formulas (3.22) through (3.25).

Table 5-6-1 presents the number of hypothetical examinees for each of the two subgroups, i.e., those who answered correctly to each of the ten unknown, binary test items and those who did not, respectively. There are seven hypothetical examinees to be excluded in Degree 4 Case, when we use the maximum likelihood estimates based upon the original Old Test and either Two-Parameter Beta Method or Normal Approach Method, which will be introduced in Sections V.7 and V.8 . For one of them, the estimated density function, $\hat{g}(\hat{\theta})$, assumes a negative value, and, for the other six, the estimated conditional variance, Var.$(\theta|\hat{\theta}_s)$, turned out to be negative. The frequencies to be subtracted from those for the success and failure groups for each of the ten unknown, binary test items are shown in brackets in Table 5-6-1. Exclusions of

TABLE 5-6-1

Numbers of Hypothetical Examinees Who Belong to the Success and Failure Groups of Each of the Ten Unknown, Binary Test Items. Negative Number Shown in Brackets After Each Entry Indicates the Number of Examinees to Be Subtracted When We Use Degree 4 Case for the Total Set of Maximum Likelihood Estimates Which Are Based Upon the Original Old Test.

Item h	Failure Subgroup	Success Subgroup
1	22 (-3)	478 (-4)
2	68 (-1)	432 (-6)
3	100 (-3)	400 (-4)
4	150 (-3)	350 (-4)
5	202 (-3)	298 (-4)
6	246 (-3)	254 (-4)
7	302 (-3)	198 (-4)
8	345 (-3)	155 (-4)
9	399 (-3)	101 (-4)
10	429 (-4)	71 (-3)

examinees happened in some other situations where we used different methods and/or different Old Tests, but the number of examinees excluded does not exceed nineteen.

In most of our studies both Degree 3 and 4 Cases were used, and sometimes Degree 5 Case was added. As it turned out, in all situations, the resultant estimated item characteristic functions of the ten unknown, binary test items are practically identical across the cases for the meaningful range of ability θ . This proves the robustness of our methods and approaches over the approximation to the density function, $g(\hat{\theta})$.

(V.7) Pearson System Method

We shall assume that the square root of the test information function, $[I(\theta)]^{1/2}$, of our Old Test is not constant, as is the case with most practical situations. Thus we need the transformation of θ to τ , and, at the end of the whole process, the retransformation of τ to θ , the rationale and actual procedure of which were described in Sections III.8 and V.3 . If the Old Test has a constant amount of test information, as is the case with our original Old Test, the reader may simply replace τ by θ . Let $\phi(\tau|\hat{\tau})$ denote the conditional density function of τ , given its maximum likelihood estimate, $\hat{\tau}$. It should be recalled that $\hat{\tau}$ is obtained from $\hat{\theta}$ through the same polynomial transformation which was introduced in Section V.3 . We can write for the first through fourth conditional moments of τ , given $\hat{\tau}$,

$$(5.20) \qquad E(\tau|\hat{\tau}) = \hat{\tau} + C^{-2}\frac{d}{d\hat{\tau}}\log g(\hat{\tau}) \quad .$$

$$(5.21) \qquad \mathrm{Var.}(\tau|\hat{\tau}) = C^{-2}[1 + C^{-2}\frac{d^2}{d\hat{\tau}^2}\log g(\hat{\tau})] \quad .$$

$$(5.22) \qquad E[\{\tau - E(\tau|\hat{\tau})\}^3|\hat{\tau}] = C^{-6}[\frac{d^3}{d\hat{\tau}^3}\log g(\hat{\tau})] \quad .$$

and

$$(5.23) \qquad E[\{\tau-E(\tau|\hat{\tau})\}^4|\hat{\tau}] = C^{-4}[3 + 6C^{-2}\{ \frac{d^2}{d\hat{\tau}^2} \log g(\hat{\tau}) + 3C^{-4}\{ \frac{d^2}{d\hat{\tau}^2} \log g(\hat{\tau})\}^2 + C^{-4}\{ \frac{d^4}{d\hat{\tau}^4} \log g(\hat{\tau})\}] \ ,$$

where C is the target constant for the square root of the test information function, $[I^*(\tau)]^{1/2}$. Substituting (5.21), (5.22) and (5.23) for μ_2 , μ_3 and μ_4 in (4.2) and (4.3), we obtain the two indices, β_1 and β_2 , and, from these two values and (4.1), Pearson's criterion κ is obtained. These indices, which can be computed for any fixed value of $\hat{\tau}$, will indicate which type of distribution of Pearson's system (Elderton and Johnson, 1969; Johnson and Kotz, 1970) we should turn to for $\phi(\tau|\hat{\tau})$. A brief summary of this procedure can be described as follows.

Type I (Beta distribution, general)	: $\kappa<0$
Type II (Beta distribution, symmetric)	: $\kappa=0$, $\beta_1=0$, $\beta_2<3$
Type III (gamma distribution)	: $\kappa=\infty$, $2\beta_2-3\beta_1-6=0$
Type IV	: $0<\kappa<1$
Type V	: $\kappa=1$
Type VI	: $\kappa>1$
Type VII (including t-distribution)	: $\kappa=0$, $\beta_1=0$, $\beta_3>3$
Normal distribution	: $\kappa=0$, $\beta_1=0$, $\beta_2=3$

The estimated conditional density function, $\hat{\phi}(\tau|\hat{\tau})$, thus approximated, has an important role in all of our four different approaches, which will be introduced in Sections V.10 through V.14 .

It is a characteristic of the Pearson System Method that we use all of the first four conditional moments of τ , given $\hat{\tau}$. Using these four conditional moments, the indices, β_1 , β_2 and κ , are obtained, and they direct us to one of the Pearson System distributions. For example, when we approximate the density function $g(\hat{\theta})$, which is based upon the original Old Test, for the total group of examinees, in Degree 3 Case, $\hat{\phi}(\theta|\hat{\theta})$ turned out to be of Type I for 318 values of $\hat{\theta}_s$, of the normal distribution for 181 values of $\hat{\theta}_s$, and for the other one case it is undefined because of the negative value for the estimated fourth conditional moment; in Degree 4 Case, $\hat{\phi}(\theta|\hat{\theta})$ proved to be of Type I for 432 values of $\hat{\theta}_s$, of Type II for 54 values of $\hat{\theta}_s$, undefined for 13 values of $\hat{\theta}_s$ because of the negative values for the estimated second and/or fourth conditional moments, and for the other one case the estimated density, $\hat{g}(\hat{\theta}_s)$, is negative and, therefore, it is undefined. If, for instance, $\hat{\phi}(\theta|\hat{\theta})$ is of Type I, then the four parameters of the Beta distribution will be estimated from the four conditional moments of θ, given $\hat{\theta}_s$, and so on.

In comparison to the other two methods, i.e., Two-Parameter Beta Method and Normal Approach Method, which will be introduced in Section V.8 and V.9, we can say Pearson System Method is theoretically sound. It will provide us with varieties of unrestricted curves for the estimated conditional density functions, $\hat{\phi}(\tau|\hat{\tau})$, which will enable us to approximate the true conditional density functions well. Its disadvantage lies in the fact that the use of higher conditional moments, like the fourth moment, may lead us to inaccuracy of estimation, as is implied in the two examples given in the preceding paragraph. If this is the case, we may use either Two-Parameter Beta Method or Normal Approach Method, which requires only the first two conditional moments.

(V.8) Two-Parameter Beta Method

Beta distribution is known for its abundance of different shapes in its density function. They include unimodal, symmetric curves, unimodal, asymmetric curves, J-shape curves, U-shape curves, and linear functions. For this reason, the distribution has been used by many researchers in approximating empirical distributions. In the Pearson System Method, which was introduced in the preceding section, Beta distribution is used as two of the Pearson System distributions, i.e., Types I and II. When we approximate the conditional density, $\hat{\phi}(\tau|\hat{\tau})$, by a Beta density function, we can write

$$(5.24) \qquad \hat{\phi}(\tau|\hat{\tau}) = [B(p_{\hat{\tau}},q_{\hat{\tau}})]^{-1}(\tau-a_{\hat{\tau}})^{p_{\hat{\tau}}-1}(b_{\hat{\tau}}-\tau)^{q_{\hat{\tau}}-1}(b_{\hat{\tau}}-a_{\hat{\tau}})^{-(p_{\hat{\tau}}+q_{\hat{\tau}}-1)},$$

where $p_{\hat\tau}$, $q_{\hat\tau}$, $a_{\hat\tau}$ and $b_{\hat\tau}$ are the four parameters of the Beta distribution, and $B(p_{\hat\tau},q_{\hat\tau})$ is the Beta function which is given by

(5.25) $$B(p_{\hat\tau},q_{\hat\tau}) = \int_0^1 u^{p_{\hat\tau}-1} (1-u)^{q_{\hat\tau}-1} du \quad .$$

These four parameters are estimated from the first four conditional moments of τ , given $\hat\tau$, and the resultant β_1 and β_2 (cf. Section IV.1). We can write

(5.26) $$\hat p_{\hat\tau} , \hat q_{\hat\tau} = (r/2)[1 \pm (r+2)\{\beta_1[\beta_1(r+2)^2 + 16(r+1)]^{-1}\}^{1/2}] \quad ,$$

(5.27) $$\hat b_{\hat\tau} - \hat a_{\hat\tau} = \{E[(\tau - E[\tau|\hat\tau])^2|\hat\tau]\}^{1/2}\{\beta_1(r+2)^2 + 16(r+1)\}^{1/2} /2 \quad ,$$

(5.28) $$\hat a_{\hat\tau} = E[\tau|\hat\tau] - \hat p_{\hat\tau}(\hat b_{\hat\tau}-\hat a_{\hat\tau})/r \quad ,$$

and

(5.29) $$\hat b_{\hat\tau} = E[\tau|\hat\tau] + \hat q_{\hat\tau}(\hat b_{\hat\tau}-\hat a_{\hat\tau})/r \quad ,$$

where

(5.30) $$r = 6(\beta_2-\beta_1-1) / (6+3\beta_1-2\beta_2) \quad .$$

When the two parameters, $p_{\hat\tau}$ and $q_{\hat\tau}$, are equal, the Beta distribution becomes Pearson's Type II distribution, and we have

(5.31) $$\hat p_{\hat\tau} = \hat q_{\hat\tau} = r/2 \quad .$$

and, otherwise, it is Pearson's Type I distribution.

When the two of the four parameters of the Beta distribution, $a_{\hat\tau}$ and $b_{\hat\tau}$, which are the lower and the upper endpoints of the interval for which the density function assumes positive values, are a priori given, the estimation of the other two parameters is much more simplified. In fact, we only need the first two conditional moments of τ , given $\hat\tau$, in addition to the set values for $a_{\hat{\tau}}$ and $b_{\hat{\tau}}$. We have

$$(5.32) \qquad p = M_1^2 \, (1-M_1) \, M_2^{-1} - M_1 \quad ,$$

and

$$(5.33) \qquad q = M_1 \, (1-M_1)^2 M_2^{-1} - (1-M_1) \quad ,$$

where M_1 and M_2 are defined by

$$(5.34) \qquad M_1 = [E(\tau|\hat{\tau})-a_{\hat{\tau}}](b_{\hat{\tau}}-a_{\hat{\tau}})^{-1} \quad ,$$

and

$$(5.35) \qquad M_2 = \mathrm{Var.}(\tau|\hat{\tau}) \, (b_{\hat{\tau}}-a_{\hat{\tau}})^{-2} \quad .$$

In the Two-Parameter Beta Method, we adopt a priori set parameters, $a_{\hat{\tau}}$ and $b_{\hat{\tau}}$, and estimate the other two parameters, $p_{\hat{\tau}}$ and $q_{\hat{\tau}}$, accordingly, and use them in (5.24) for the estimated conditional density, $\hat{\phi}(\tau|\hat{\tau})$. It has an advantage over the Pearson System Method in the sense that we only need the first two conditional moments of τ , given $\hat{\tau}$, instead of four, and yet we can make use of the abundance of different shapes of the Beta density function. The biggest problem is how to select suitable values for $a_{\hat{\tau}}$ and $b_{\hat{\tau}}$ for each fixed value of $\hat{\tau}$. In the present study, these values are chosen relatively arbitrarily, and we adopted

$$(5.36) \qquad \begin{cases} a_{\hat{\tau}} = \hat{\tau} - 2.55C^{-1} \\ b_{\hat{\tau}} = \hat{\tau} + 2.55C^{-1} \end{cases} \quad ,$$

where C is the target constant square root of the test information function, $[I^*(\tau)]^{-1}$. Actually, this method was used only for the original Old Test, so C^{-1} equals σ (= 0.215) .

Although all the results obtained in the present study turned out to be as good as they can be, which will be introduced in later sections, the selection of suitable values for $a_{\hat{\tau}}$ and $b_{\hat{\tau}}$ is yet to be investigated in future, to make Two-Parameter Beta Method theoretically sounder and more useful.

(V.9) Normal Approach Method

A simple, straightforward method of approximating the conditional density function, $\phi(\tau|\hat{\tau})$, using the only first two conditional moments of τ, given $\hat{\tau}$, may be the approximation by a normal density function. We can write

$$(5.37) \qquad \hat{\phi}(\hat{\tau}|\hat{\tau}) = [2\pi\mu_2]^{-1/2} \exp[-(\tau-\mu_1')^2 (2\mu_2)^{-1}] \quad ,$$

where

$$(5.38) \qquad \mu_1' = E[\tau|\hat{\tau}] \quad ,$$

and

$$(5.39) \qquad \mu_2 = \mathrm{Var.}[\tau|\hat{\tau}] \quad .$$

An advantage of this method over the Pearson System Method is that we need only the first two conditional moments, and one over the Two-Parameter Beta Method is that we do not need any a priori set parameters. A disadvantage is obviously that it restricts the estimated conditional moment to be a unimodal, symmetric function, regardless of its true shape. In spite of this restriction, however, Normal Approach Method worked very well both in combination with the Bivariate P.D.F. Approach and with the Conditional P.D.F. Approach, the results of which we shall see in succeeding sections.

(V.10) Bivariate P.D.F. Approach

As was introduced in Section V.5, in the Normal Approximation Method, we approximate the bivariate distribution of τ and $\hat{\tau}$, or θ and $\hat{\theta}$ if the square root of the test information function of our Old Test is constant, by a bivariate normal distribution, for each subpopulation of examinees who share the same item score to item h. Our results turned out to be quite successful.

Figure 5-10-1 presents the theoretical regression of θ on $\hat{\theta}$ which is based upon the original Old Test, and the intervals of the standard error above and below this regression, by dotted lines,

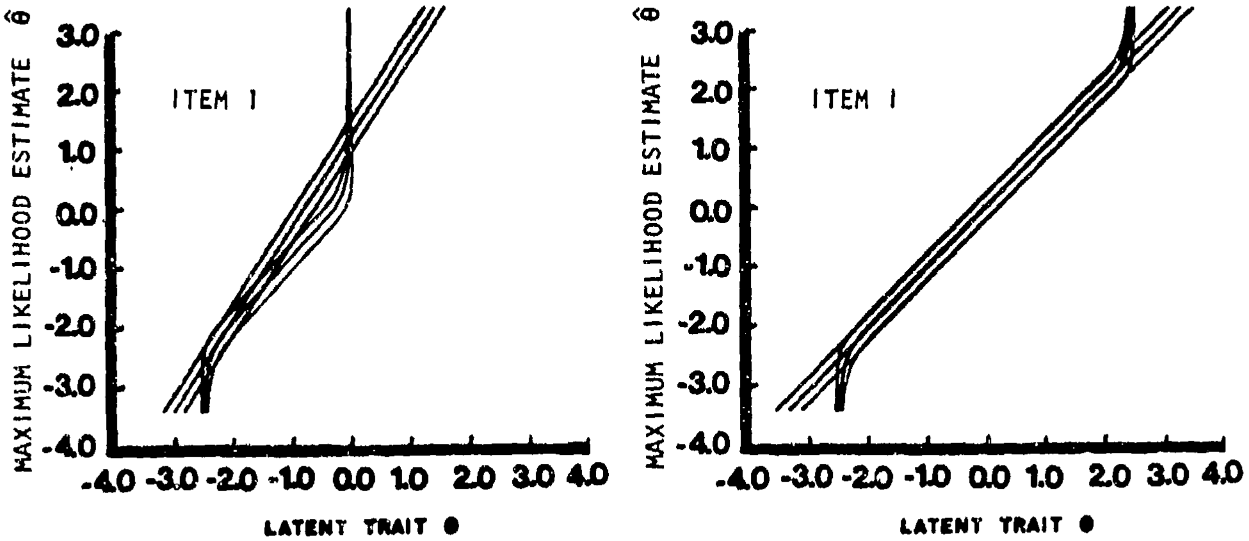


FIGURE 5-10-1

Comparison of the Theoretical Regression for Ability θ on Its Maximum Likelihood Estimate $\hat{\theta}$ (Dotted Line) with the Best Fitted Line of Ability θ, on $\hat{\theta}$ (Dashed Line), for Each Item Score Group of Item 1. Also the Standard Errors of Estimation Are Shown on Each Side of the Regression, and of the Best Fitted Line.

for each of the success and the failure groups for item 1. In the same figure also presented by dashed lines are the empirical linear regression of θ on $\hat{\theta}$, with the intervals of the empirical standard error above and below the linear regression, which were introduced in Section V.5. We can see in this figure that for the success group these two sets of curves are almost identical for most of the meaningful range of $\hat{\theta}$, while the discrepancies are substantial for the failure group. This example of the failure group for item 1 is the only extreme case, and, in fact, thirteen out of the remaining eighteen cases provided us with similar results as the one for the success group for item 1 , four cases show slight discrepancies, and the other one case lies somewhere between the two examples in Figure 5-10-1 in diversion.

The results illustrated in Figure 5-10-1 suggest that we may need to investigate some other approach than the approximation by the bivariate normal distribution to the joint distribution of τ and $\hat{\tau}$. This can be done by making use of the marginal density functions of $\hat{\tau}$ and the conditional density functions of τ , given $\hat{\tau}$, for the separate subpopulation of examinees.

Let $g_{x_h}(\hat{\tau})$ denote the proportion of the density function of the maximum likelihood estimate $\hat{\tau}$ for the subpopulation of the examinees who share the same item score, x_h $(=0,1,\ldots,m_h)$, and $\phi_{x_h}(\tau|\hat{\tau})$ and $\xi_{x_h}(\tau,\hat{\tau})$ be the corresponding conditional density of τ , given $\hat{\tau}$, and the proportion of the bivariate density of τ and $\hat{\tau}$, respectively. We can write

$$(5.40) \qquad \xi_{x_h}(\tau,\hat{\tau}) = \phi_{x_h}(\tau|\hat{\tau})\, g_{x_h}(\hat{\tau}) \quad ,$$

where

$$(5.41) \qquad g(\hat{\tau}) = \sum_{x_h=0}^{m_h} g_{x_h}(\hat{\tau})$$

and

$$(5.42) \qquad \xi(\tau,\hat{\tau}) = \sum_{x_h=0}^{m_h} \xi_{x_h}(\tau,\hat{\tau}) \quad .$$

To obtain the estimate of the proportion of the bivariate density, $\xi_{x_h}(\tau,\hat{\tau})$, we classify the set of N $\hat{\tau}_i$'s into (m_h+1)

item score categories, depending upon the item score x_h $(=0,1,\ldots,m_h)$ the examinee i obtained for a new test item h, for which the operating characteristics are to be estimated. The method of moments is applied for each of these (m_h+1) subsets of $\hat{\tau}$, and the shared density function, $g_{x_h}(\hat{\tau})$, is estimated for each subgroup. The conditional moments of τ, given $\hat{\tau}$, are also obtained for separate subgroups, using the formulas (5.20) through (5.23), with the replacement of $g(\hat{\tau})$ by $(N/N_{x_h})\, g_{x_h}(\hat{\tau})$, where N_{x_h} denotes the number of examinees whose item scores to item h are x_h. Based on these estimated conditional moments, the parameters of a specific density function, which is adopted for $\phi_{x_h}(\tau|\hat{\tau})$, are obtained for each subgroup x_h. The choice of $\phi_{x_h}(\tau|\hat{\tau})$ depends upon which of the three methods, i.e., Normal Approach Method, Two-Parameter Beta Method and Pearson-System Method, is taken. The bivariate density function of $\hat{\tau}$ and τ is obtained from (5.40) for each of the (m_h+1) subgroups. Then the estimated operating characteristic, $\hat{p}_{x_h}(\theta)$ $[= p^*_{x_h}(\tau(\theta))]$, is given by

$$(5.43) \qquad \hat{P}_{x_h}(\theta) = \int_{-\infty}^{\infty} \hat{\xi}_{x_h}(\hat{\tau},\tau)d\hat{\tau} \; [\sum_{j=0}^{m_h} \int_{-\infty}^{\infty} \hat{\xi}_j(\hat{\tau},\tau)d\hat{\tau}]^{-1} \;, \qquad x_h = 0,1,\ldots,m_h \;.$$

This approach was applied to our data (cf. Section III.3) in combination with the Normal Approach Method (cf. Section V.9) for Degree 3, 4 and 5 Cases. We used the five hundred maximum likelihood estimates, $\hat{\theta}_s$, which were based upon the original Old Test. The polynomials of degrees 3, 4 and 5 approximating $g_{x_h}(\hat{\tau})$ for each of the two subpopulations, i.e., the success and the failure groups, are illustrated for $h = 6$ in Section V.6 as Figure 5-6-1.

Figure 5-10-2 presents the resultant estimated ability distributions in Degree 3 (dotted curve), 4 (short, dashed curve) and 5 (long, dashed curve), together with the theoretical density (solid curve) and the frequency distribution of θ (histogram with

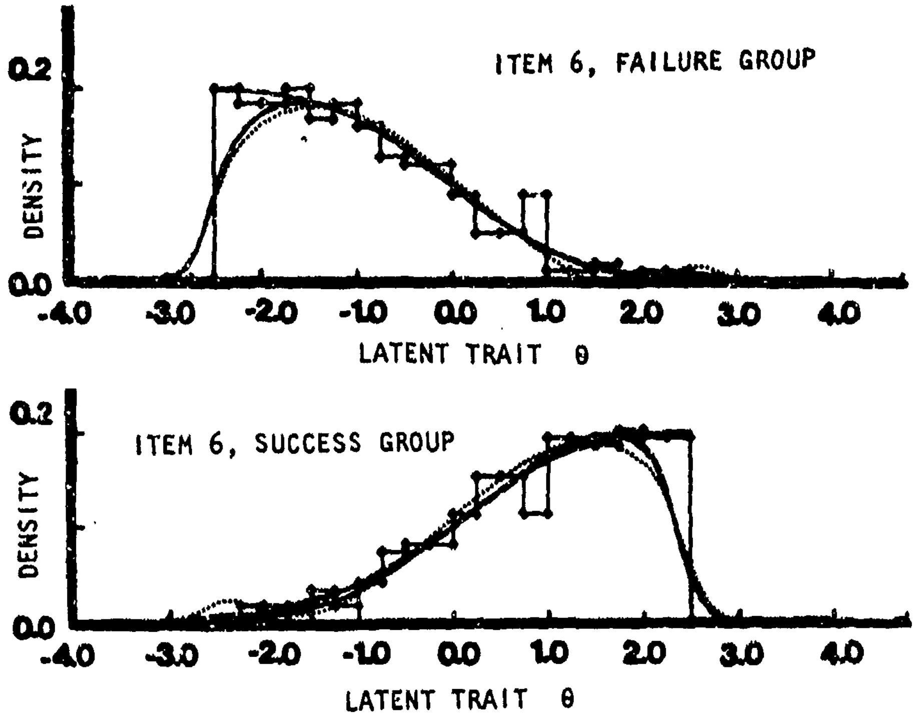


FIGURE 5-10-2

Estimated Proportions of the Density Function of Ability θ in Degree 3 (Dotted Curve), 4 (Short, Dashed Curve) and 5 (Long, Dashed Curve) Cases of the Bivariate P.D.F. Approach with the Normal Approach Method, for Each of the Success and Failure Subpopulations. Actual Frequencies (Solid Line with Diamonds) and the Theoretical Proportion of the Density Function (Solid Curve) Are Also Drawn.

solid diamonds), for each of the success and failure groups. We can see in this figure that, except for the lower end of θ for the failure group and the upper end of θ for the success group, these three curves of Degree 3, 4 and 5 Cases are very close to the theoretical curves. The results for the other nine binary test items are similar to this example. In some cases the fit is best in Degree 5 Case and worst in Degree 3 Case, but this order is not true with all the cases. In most cases, the resultant three curves are close to one another, as we can see in Figure 5-10-2.

Figure 5-10-3 presents the resultant three estimated item characteristic functions of Degree 3, 4, and 5 Cases for item 6, which were obtained from (5.43) with $x_h = 1$ and $m_h = 2$, by dotted, short dashed and long dashed curves, respectively. We can see in this figure that all these results are close to the theoretical item characteristic function, which is also shown in the figure by a solid curve, and are much closer than the frequency ratios of θ for the correct answer, which are shown by a solid line with diamonds in the figure. If we compare these three estimated item characteristic functions with one another, we can say that the result of Degree 3 Case is not as good as the other two. This is not a general tendency, however. For most of the other nine test items, the resultant estimated item characteristic functions of Degree 3 Case are much closer to the corresponding theoretical item characteristic functions, and, in fact, for item 7 it shows the best fit among the three.

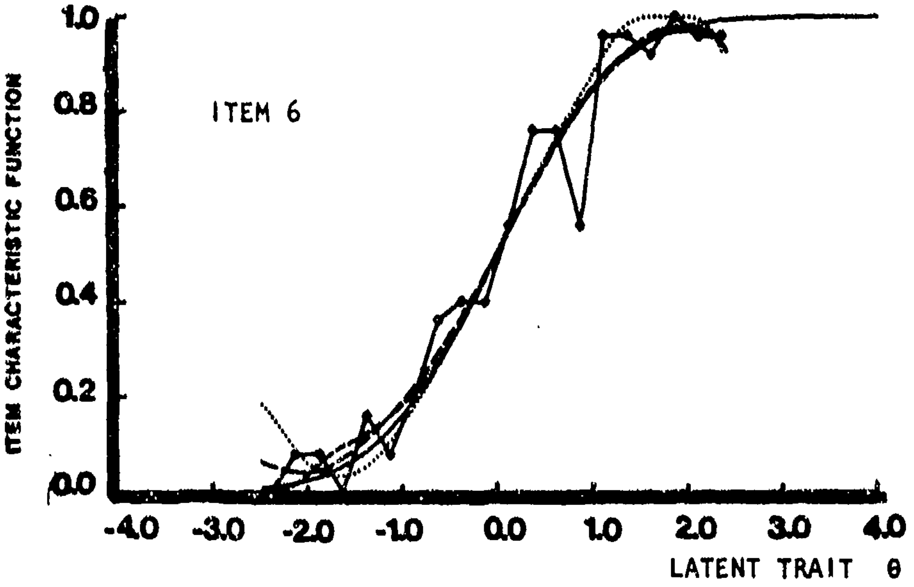


FIGURE 5-10-3

Estimated Item Characteristic Functions of Item 6 for Degree 3 (Dotted Curve), 4 (Short, Dashed Curve) and 5 (Long, Dashed Curve) Cases of the Bivariate P.D.F. Approach with the Normal Approach Method, Together with the Theoretical Item Characteristic Function (Solid Curve) and the Actual Frequency Ratios (Solid Line with Diamonds).

(V.11) Histogram Ratio Approach

In this approach, and also in the Curve Fitting Approach and the Conditional P.D.F. Approach, which will be introduced in the following two sections, we make use of the estimated conditional density function of τ , which is evaluated for the maximum likelihood estimate, $\hat{\tau}_s$, of each individual examinee s . This is the difference of these three approaches from the Bivariate P.D.F. Approach, in which $\hat{\phi}(\tau|\hat{\tau})$ is used for approximating the bivariate density function, $\xi(\tau,\hat{\tau})$, as we have observed in the preceding section.

Using the Monte Carlo method, we have the computer produce a specified number of τ following the estimated conditional density function, $\hat{\phi}(\tau|\hat{\tau}_s)$, for each value of $\hat{\tau}_s$. Let $\tilde{\tau}$ denote the values of τ thus produced, as we did in the Normal Approximation Method, and ν be the number of $\tilde{\tau}$'s produced for each $\hat{\tau}_s$. The resultant set of $\tilde{\tau}$'s are classified into (m_h+1) categories, depending upon the item score x_h $(=0,1,\ldots,m_h)$ which the examinee s obtained for item h . Then each $\tilde{\tau}$ is transformed to $\hat{\theta}$, by means of

$$(5.44) \qquad \theta = \tau^{-1}[\tau(\theta)] \quad .$$

When $\tau(\ )$ is given by the polynomial shown as (5.14), this process can easily be performed by the Newton-Raphson Method.

We divide the interval of θ of our interest into subintervals of equal width. Let t denote the subinterval, θ_t be the midpoint of the subinterval t , and $H_{x_h}(\tilde{\theta}\varepsilon t)$ denote the frequency of $\tilde{\theta}$'s , which belong to the item score x_h and the subinterval t . We have for the estimated operating characteristic of the item score x_h

$$(5.45) \qquad \hat{P}_{x_h}(\theta_t) = H_{x_h}(\tilde{\theta}\varepsilon t)[\sum_{j=0}^{m_g} H_j(\tilde{\theta}\varepsilon t)]^{-1} \quad , \quad x_h=0,1,\ldots,m_h \quad .$$

This approach was applied to the set of five hundred maximum likelihood estimates $\hat{\theta}_s$, which were obtained upon the original Old Test, in combination with the Two-Parameter Beta Method for approximating the conditional density function, $\phi(\theta|\hat{\theta})$. The number of hypothetical examinees actually used in Degree 4 Case is

493 (cf. Section V.6), while in Degree 3 Case the total 500 examinees were used. In both cases, we adopted $\nu = 5$, and 0.25 for the subinterval width. Figure 5-11-1 presents the resultant estimated item characteristic functions of item 6 for Degree 3 Case by triangles, and for Degree 4 Case by squares, respectively.

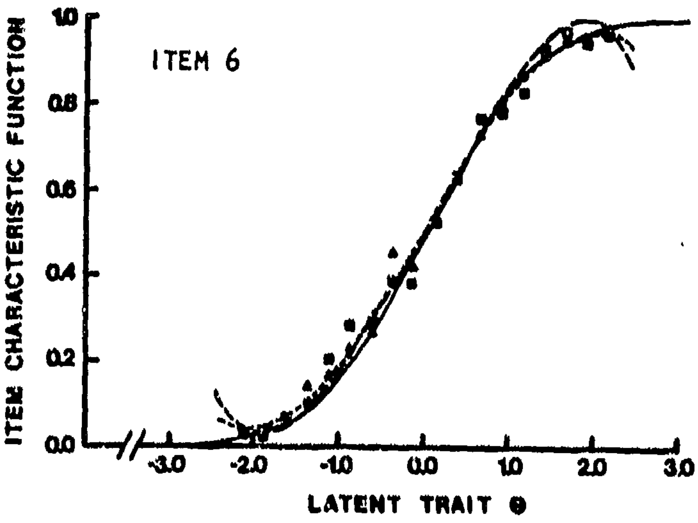


FIGURE 5-11-1

Estimated Item Characteristic Functions of Item 6 for Degree 3 (Triangles) and 4 (Squares) Cases of the Histogram Ratio Approach and Those for Degree 3-3 (Long, Dashed Curve) and 3-4 (Short, Dashed Curve) Cases of the Curve Fitting Approach, with the Two-Parameter Beta Method.

We can see that the two sets of estimates are fairly close to the theoretical item characteristic function of item 6, which is drawn by a solid curve in Figure 5-11-1. It is expected that the fitness will be even better if we increase ν , and decrease the subinterval width. Similar results were obtained for each of the other nine binary test items.

An advantage of the Histogram Ratio Approach over the others lies in its simplicity and straightforwardness. In order to obtain a smooth curve for the estimated operating characteristic, it is advisable to use a fairly large number for ν , and a small width for the subinterval, t , of θ .

(V.12) Curve Fitting Approach

This approach follows the same process as the Histogram Ratio Approach until we obtain νN $\tilde{\theta}$'s , which are divided into (m_h+1) subsets of item scores x_h for item h . Then for each subset of $\hat{\theta}$'s a polynomial of a specified degree is fitted by the method of moments. Let $\eta_{x_h}(\theta)$ denote such a polynomial fitted for the $\hat{\theta}$'s of the subset x_h . The estimated operating characteristic of the item score x_h is given by

$$(5.46) \qquad \tilde{P}_{x_h}(\theta) = \eta_{x_h}(\theta)[\sum_{j=0}^{m_h} \eta_j(\theta)]^{-1} \quad , \quad x_h=0,1,\ldots,m_h \ .$$

This approach was applied to the same set of $\tilde{\theta}$'s as we obtained for the Histogram Ratio Approach in the preceding section. Both polynomials of degree 3 and degree 4 were fitted to the resultant two subsets of $\tilde{\theta}$'s , which were obtained in each of Degree 3 and Degree 4 Cases. We shall call these four cases Degree 3-3, 3-4, 4-3 and 4-4 Cases, with the second number indicating the degree of the polynomials fitted to the subsets of $\tilde{\theta}$'s . An example of the curve fitting for Degree 3-3 and 3-4 Cases for item 4 was given in Section IV.1 as Figure 4-1-3.

The resultant estimated item characteristic function for item 6 in Degree 3-3 and 3-4 Cases are shown in Figure 5-11-1 in the preceding section by long and short dashes, respectively, together with the results obtained by the Histogram Ratio Approach. Figure 5-12-1 presents the corresponding results for Degree 4-3 and 4-4 Cases by long and short dashes, respectively. We can see that all of these four results are very close to the theoretical item characteristic function, except for both ends of the curves. In this example of item 6, we can say the curve for Degree 3-4 fits the best to the theoretical item characteristic function. We cannot generalize this to the other items, however, and there is no systematic tendencies as to which of the four cases provides us with best fitting curves.

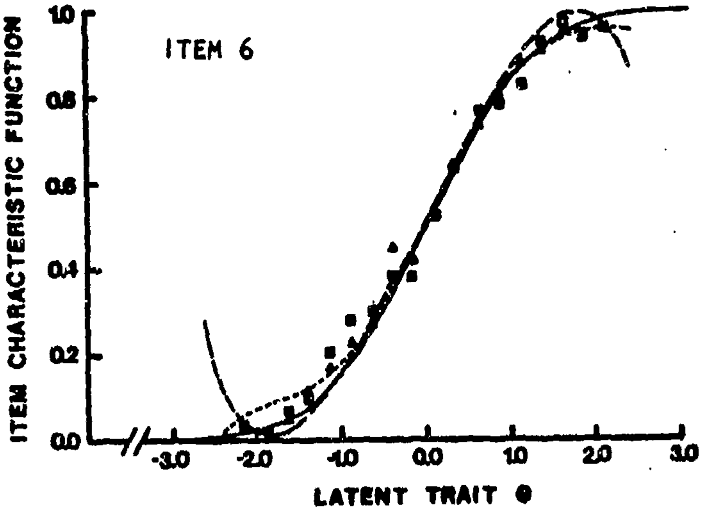


FIGURE 5-12-1

Estimated Item Characteristic Functions of Item 6 for Degree 3 (Triangles) and 4 (Squares) Cases of the Histogram Ratio Approach and Those for Degree 4-3 (Long, Dashed Curve) and 4-4 (Short, Dashed Curve) Cases of the Curve Fitting Approach, with the Two-Parameter Beta Method.

(V.13) Conditional P.D.F. Approach

In this approach, we use the whole approximation to the conditional density function, $\hat{\phi}(\tau|\hat{\tau}_s)$. In the Simple Sum Procedure, we have for the operating characteristic of the item score x_h

$$(5.47) \qquad \hat{P}_{x_h}(\theta) = \hat{P}^*_{x_h}[\tau(\theta)] = \sum_{s\epsilon x_h} \hat{\phi}(\tau|\hat{\tau}_s)[\sum_{s=1}^{N} \hat{\phi}(\tau|\hat{\tau}_s)]^{-1} \quad , \qquad x_h=0,1,\ldots,m_h \quad .$$

In the present study, this approach was frequently used. Among others, it was used for the comparison of the results obtained upon several different Old Tests, which will be introduced in Chapter 6.

It should be noted that we can write for the conditional density of τ , given $\hat{\tau}_s$,

$$(5.48) \qquad \phi(\tau|\hat{\tau}_s) = \psi(\hat{\tau}_s|\tau)\ f^*(\tau)\ [\int_{-\infty}^{\infty} \psi(\hat{\tau}_s|\tau)\ f^*(\tau)\ d\tau]^{-1} \quad ,$$

where $\psi(\hat{\tau}_s|\tau)$ is the conditional density of $\hat{\tau}_s$, given τ , and $f^*(\tau)$ is the marginal density of τ . From our simulated data, we can obtain this theoretical density function, by using $n(\tau,C^{-1})$ for $\psi(\hat{\tau}_s|\tau)$, where C is the target square root of the test information function, $[I^*(\tau)]^{1/2}$, and

$$(5.49) \qquad f^*(\tau) = f(\theta)\frac{d\theta}{d\tau} = f(\theta)\ C\ [\sum_{k=0}^{m} \alpha_k\ \theta^k]^{-1}$$

$$\begin{cases} = 0.2C[\sum_{k=0}^{7} \alpha_k\ \theta^k]^{-1} & \text{for} \quad \tau(-2.5) < \tau < \tau(2.5) \\ = 0 & \text{otherwise.} \end{cases}$$

We can replace $\hat{\phi}(\tau|\hat{\tau}_s)$ in (5.47) by $\phi(\tau|\hat{\tau}_s)$ thus obtained, and the resultant function is called the criterion operating characteristic of item score x_h . This function is the limiting case that we can possibly attain by adopting the Simple Sum Procedure of the Conditional P.D.F. Approach upon a given set of data.

Figures 5-13-1 through 5-13-3 present the three sets of estimated item characteristic functions of item 6, obtained by the Conditional P.D.F. Approach, in comparison with the theoretical item characteristic function, which is drawn by a thick, solid curve, and the frequency ra*ios of the correct answer, which are shown by the combination of long dashes and dots. These resultant estimated operating characteristics are based upon the Conditional P.D.F. Approach combined with the Two-Parameter Beta Method, Normal Approach Method and Pearson System Method, respectively. In each figure, the result obtained in Degree 3 Case is plotted by long dashes, and the one obtained in Degree 4 Case is drawn by short, thick dashes, respectively. There is the fifth curve, plotted by a thin, solid curve in each figure, i.e., the criterion item characteristic function of item 6. It is hard to single it out,

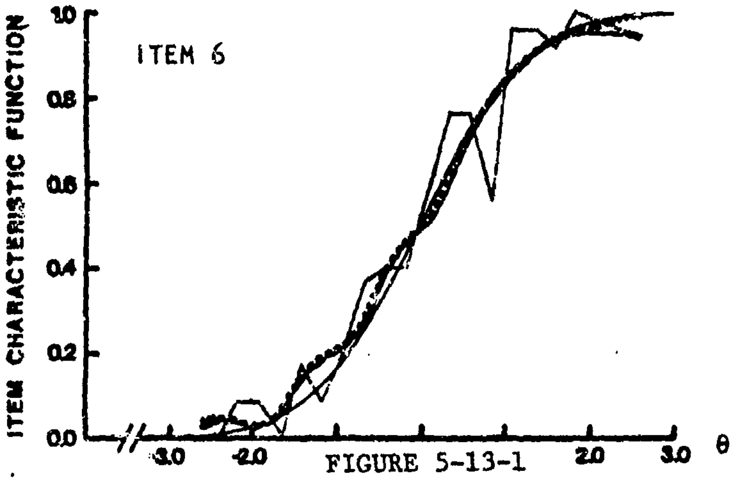


FIGURE 5-13-1

Estimated Item Characteristic Functions Obtained by the Conditional P.D.F. Approach with the Three-Parameter Beta Method, in Degree 3 (Long Dashes) and 4 (Short, Thick Dashes) Cases, in Comparison with the Criterion Item Characteristic Function (Thin, Solid Curve), the Frequency Ratios of the Correct Answer (Long Dashes and Dots), and the Theoretical Item Characteristic Function (Thick, Solid Curve).

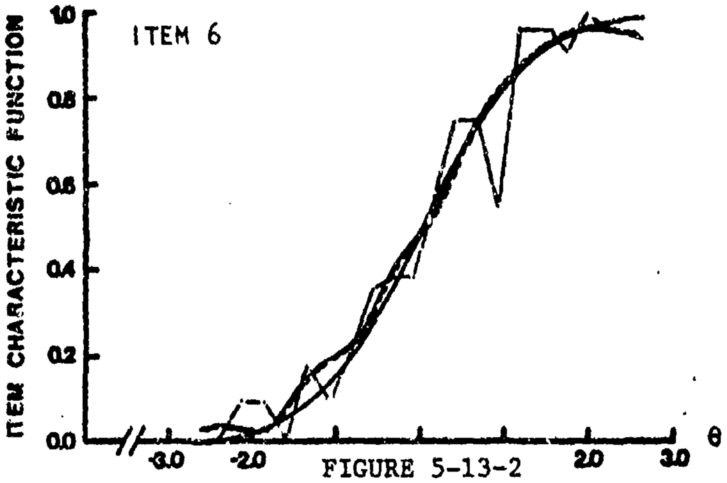


FIGURE 5-13-2

Result of the Normal Approach Method, in Comparison with the Other Three.

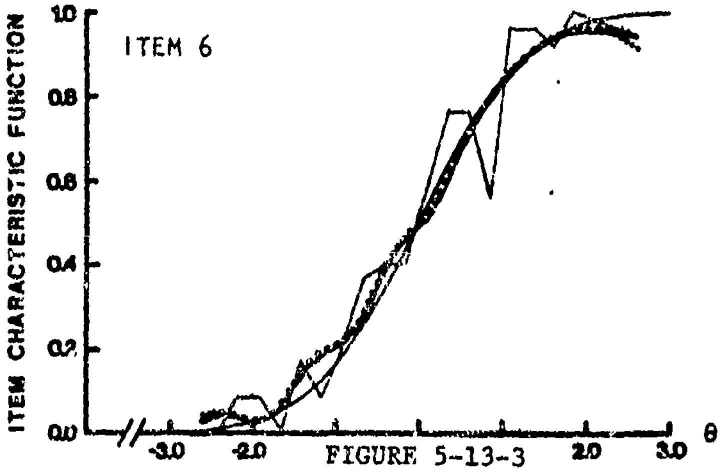


FIGURE 5-13-3

Result of the Pearson System Method, in Comparison with the Other Three.

however, because in each figure the three curves, i.e., the results of Degree 3 and 4 Cases and the criterion item characteristic function, are practically indistinguishable. This result is not unique for item 6, which we have chosen as an example more or less arbitrarily. In fact, for the interval of θ , (-2.2, 2.2) , the three curves are practically identical for each of the other nine binary test items, although outside of this interval of θ there are some discrepancies.

The above results indicate the high success of using either one of the three methods, i.e., Two-Parameter Beta Method, Normal Approach Method and Pearson System Method, in approximating the conditional density function, $\phi(\tau|\hat{\tau}_s)$. We have investigated the fitness of these curves further, some results of which are illustrated in Figures 5-13-4 through 5-13-6.

Figure 5-13-4 presents the regression, $E[\theta|\hat{\theta}]$, of ability θ on its maximum likelihood estimate $\hat{\theta}$, which is based upon

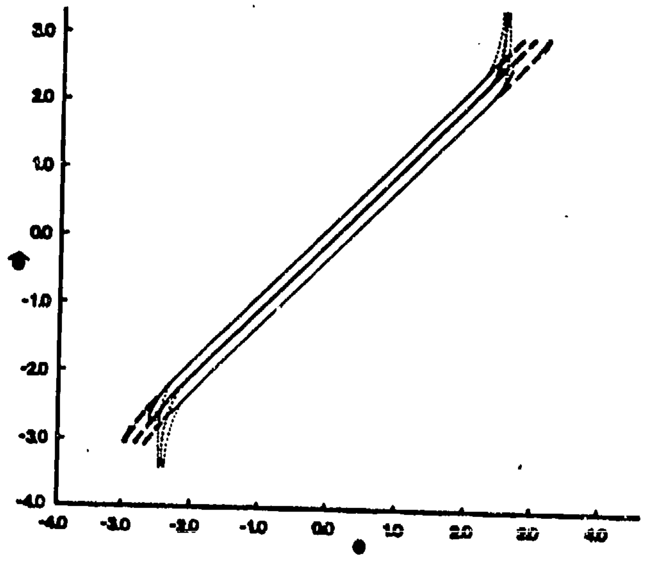


FIGURE 5-13-4

Regression of Ability θ on Its Maximum Likelihood Estimate $\hat{\theta}$ Based Upon the Original Old Test.

the original Old Test, with the intervals of the standard error, $[\text{Var.}(\theta|\hat{\theta})]^{1/2}$, on each side, by dots. These values were obtained by

$$(5.50) \qquad E(\tau|\hat{\tau}) = \int_{-\infty}^{\infty} \tau \; \phi(\tau|\hat{\tau}) \, d\tau \quad ,$$

and

$$(5.51) \qquad \text{Var.}(\tau|\hat{\tau}) = \int_{-\infty}^{\infty} [\tau - E(\tau|\hat{\tau})]^2 \; \phi(\tau|\hat{\tau}) \, d\tau \quad ,$$

where $\phi(\tau|\hat{\tau})$ is defined by (5.48), with the replacement of τ by θ, and $\hat{\tau}_s$ by $\hat{\theta}$. In the same figure, also presented by dashed and solid lines are the corresponding estimates in Degree 3 and 4 Cases. We can see that these three sets of curves are practically identical for the interval of θ, (-2.2, 2.0), and then divert from one another outside of this interval. This result proves a high accuracy in the estimation of the first and the second conditional moments of θ, given $\hat{\theta}$, which was done by (3.22) and (3.23), using the polynomial obtained by the method of moments as the estimated density function, $\hat{g}(\hat{\theta})$, in both Degree 3 and 4 Cases. The differences between the two cases in the diversion from the true regression outside of the interval of $\hat{\theta}$, (-2.2, 2.0), are due to the differences between the two polynomials around these two areas, which are shown in Figure 4-1-2.

From the result shown in Figure 5-13-4, we can expect that the fitness of $\hat{\phi}(\theta|\hat{\theta}_s)$ to $\phi(\theta|\hat{\theta}_s)$ should be better for the interval of $\hat{\theta}$, (-2.2, 2.0), than for the range of $\hat{\theta}$ outside of this interval. Figures 5-13-5 and 5-13-6 present two examples of the fitnesses of the estimated conditional density functions to the true density function, $\phi(\theta|\hat{\theta}_s)$. These two sets of results are for $s = 50$ and $s = 500$, whose maximum likelihood estimates, $\hat{\theta}_s$, are -0.0066 and 2.6346, respectively. In both figures, the theoretical density, $\phi(\theta|\hat{\theta}_s)$, is drawn by a solid curve, and the estimated density functions, $\hat{\phi}(\theta|\hat{\theta}_s)$, obtained by the Normal

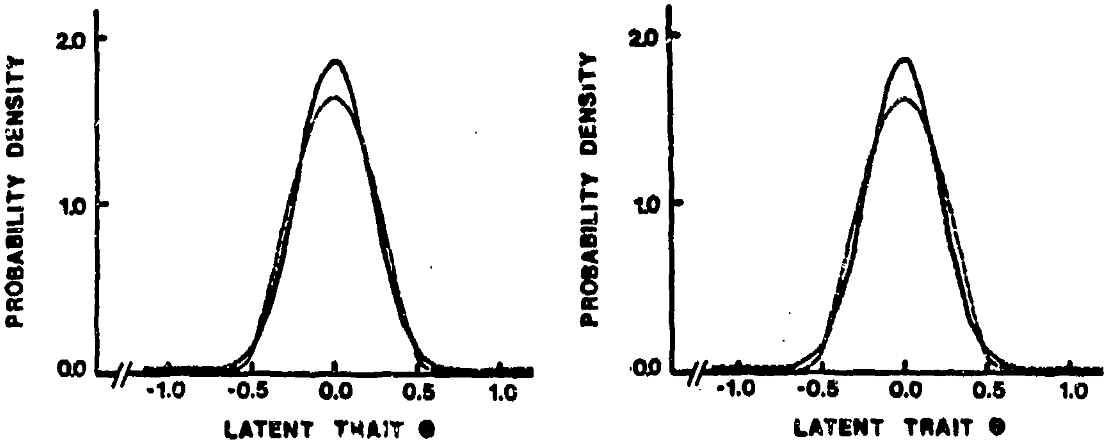


FIGURE 5-13-5

Conditional Density of θ , Given $\hat{\theta}$ (Solid Curve) and Its Estimates by the Normal Approach Method (Dotted Curve) and by the Two-Parameter Beta Method (Long Dashed Curve), for Degree 3 Case (Left) and Degree 4 Case (Right), Based upon the Original Old Test. $\hat{\theta} = \hat{\theta}_{50} = -0.0066$.

Approach Method and the Two-Parameter Beta Method, are plotted by short and long dashes, respectively, in each of the Degree 3 and 4 Cases. In Figure 5-13-5, we can see that $\hat{\phi}(\theta|\hat{\theta}_s)$, which is obtained by the Normal Approach Method, is practically identical with the theoretical density function, while the one obtained by the Two-Parameter Beta Method is somewhat different, in each of Degree 3 and 4 Cases. In this example, Pearson System Method directs us to the normal distribution in Degree 3 Case, and to the Type II Beta distribution in Degree 4 Case. The normal density curve in the left hand side graph of Figure 5-13-5, therefore, is also the result obtained by the Pearson System Method, and the one in the right hand side graph is practically identical with the one obtained by the Pearson System Method ($\beta_2 = 2.999$) . We can also see in Figure 5-13-5 that the two sets of results obtained for

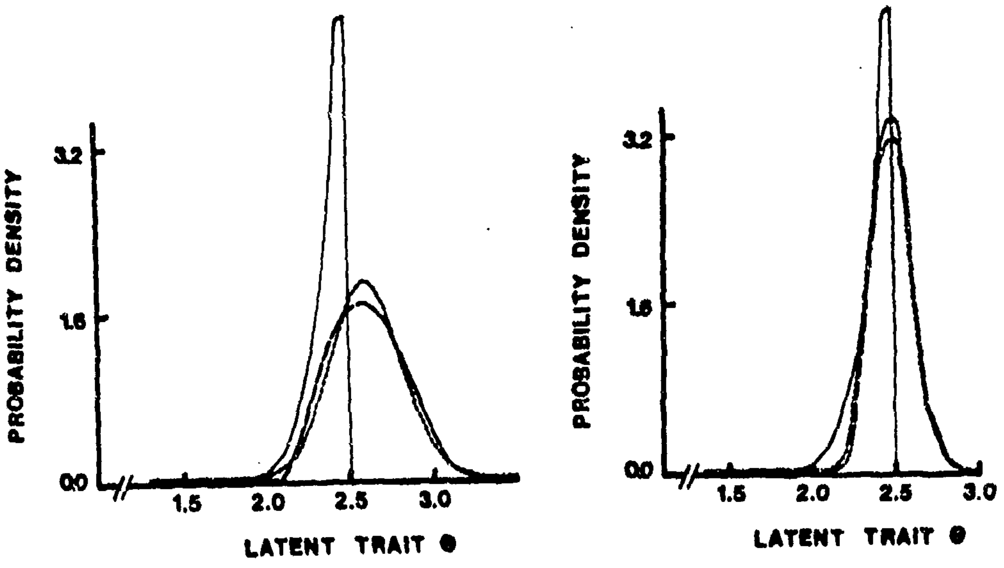


FIGURE 5-13-6

Conditional Density of θ , Given $\hat{\theta}$ (Solid Curve) and Its Estimates by the Normal Approach Method (Dotted Curve) and by the Two-Parameter Beta Method (Long Dashed Curve), for Degree 3 Case (Left) and Degree 4 Case (Right), Based upon the Original Old Test. $\hat{\theta} = \hat{\theta}_{500} = 2.6346$.

Degree 3 and 4 Cases are very close to each other.

In contrast to this, Figure 5-13-6 shows lower degrees of fitness of $\hat{\phi}(\theta|\hat{\theta}_s)$ to its theoretical counterpart, $\phi(\theta|\hat{\theta}_s)$, in both Degree 3 and 4 Cases. The departure from the theoretical density function is greater for Degree 3 Case in both results obtained by the Two-Parameter Beta Method and Normal Approach Method, which is anticipated from the greater diversion of the estimated regression of θ on $\hat{\theta}$ from the true regression in Degree 3 Case, as we have seen in Figure 5-13-4. Pearson System Method directs us to the Type I Beta distribution ($\kappa = -0.010$, $\beta_1 = 0.000$, $\beta_2 = 2.990$) in Degree 3 Case, and the distribution is undefined in Degree 4 Case.

We have sampled 42 examinees out of 493, and observed the fitnesses of the estimated density functions to the true ones (cf.

RR-78-2). As is expected from Figure 5-13-4, in most cases the results turned out to be similar to the one for $s = 50$, which we have seen in Figure 5-13-5, and in a few cases in which $\hat{\theta}_s$ lies outside of the interval, $(-2.2, 2.0)$, the results were similar to the one for $s = 500$, which we have observed in Figure 5-13-6.

Weighted Sum Procedure is an expansion of the Simple Sum Procedure, in which the estimated operating characteristic, $\hat{P}_{x_h}(\theta)$, of the item response x_h can be written as

$$(5.52)\qquad \hat{P}_{x_h} = \sum_{s\varepsilon x_h} w(\hat{\tau}_s)\,\hat{\phi}(\tau|\hat{\tau}_s)\Big[\sum_{s=1}^{N} w(\hat{\tau}_s)\hat{\phi}(\tau|\hat{\tau}_s)\Big]^{-1} ,$$

$$x_h = 0,1,\ldots,m_h \quad ,$$

where $w(\hat{\tau}_s)$ is an appropriate weight assigned to the maximum likelihood estimate $\hat{\tau}_s$ for the individual examinees. Simple Sum Procedure can be considered, therefore, as a special case of the Weighted Sum Procedure, in which $w(\hat{\tau}_s) = 1$ for all the individual examinees.

Figure 5-13-7 presents the estimated density functions of ability θ, which is divided into two portions for the success and failure subpopulations for item 6, respectively, as the results of the Weighted Sum Procedure of the Conditional P.D.F. Approach, which is combined with the Two-Parameter Beta Method. These results were obtained upon the original Old Test, using the area under the curve of $\hat{g}(\hat{\theta})$ for the subinterval of $\hat{\theta}$ which is taken from the midway between each $\hat{\theta}_s$ and the lower adjacent value of $\hat{\theta}_s$ and ends with the midpoint between $\hat{\theta}_s$ and the upper adjacent value of $\hat{\theta}_s$. The result of Degree 3 Case is plotted by dots and the one obtained by Dagree 4 Case is drawn by dashes, in each of the two graphs of Figure 5-13-7. In this figure, the theoretical portions of the density of ability θ are drawn by solid curves, the actual frequencies of θ by solid lines with

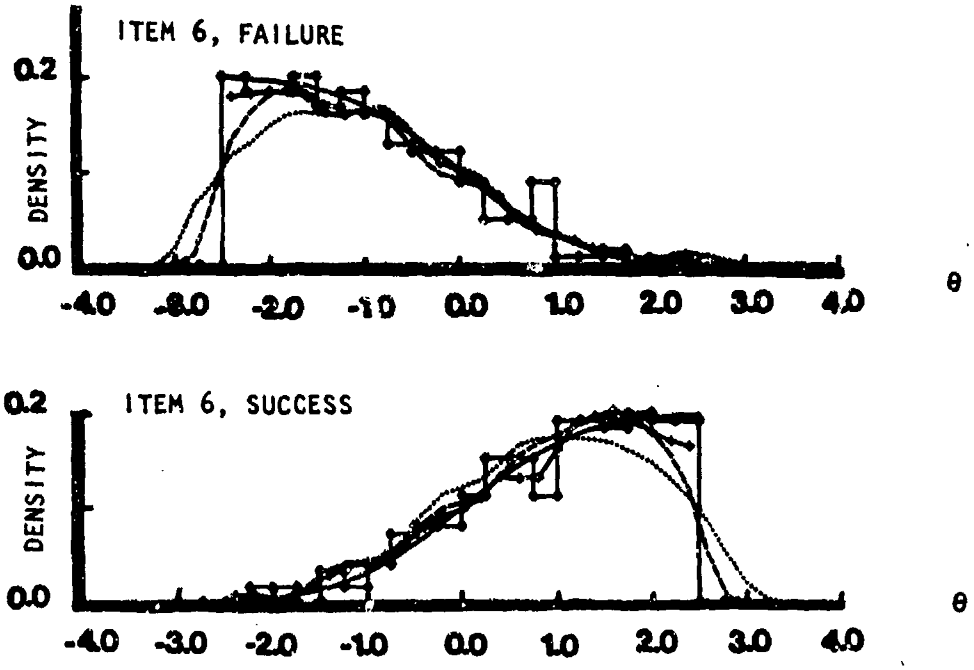


FIGURE 5-13-7

Estimated Density Functions of Ability θ Divided into Two Portions for the Success and the Failure Subpopulations for Item 6 , Obtained by the Weighted Sum Procedure of the Conditional P.D.F. Approach with the Two-Parameter Beta Method, in Degree 3 (Dotted Curves) and 4 (Dashed Curves) Cases , in Comparison with the Theoretical Portions of the Density Function (Solid Curves) , the Actual Frequencies of θ (Solid Lines with Diamonds), and the Portions for the Criterion Item Characteristic Function in the Simple Sum Procedure (Solid Curves with Crosses).

diamonds, and the functions which are the basis of the criterion item characteristic function in the Simple Sum Procedure are shown by solid curves with crosses, respectively. We can see in this result that the estimated ability distributions are more deviated from the true ability distributions in Degree 3 Case, in comparison with those of Degree 4 Case. This is not only true with item 6 but is common among the results obtained for the other nine binary test items, and also among those obtained by using the Pearson System Method instead of the Two-Parameter Beta Method. This diversion is due to the fact that we used the areas under the estimated density function, $\hat{g}(\hat{\theta})$, as the weight, $w(\hat{\theta}_s)$, and the discrepancies of $\hat{g}(\hat{\theta})$ from the true density function in Degree 3 Case are greater than the one in Degree 4 Case, as we have seen in

Figure 4-1-2.

Figures 5-13-8 and 5-13-9 present the resultant estimated item characteristic functions of item 6, in Degree 3 Case by dotted curves and in Degree 4 Case by long, dashed curves, which were obtained by the Pearson System Method and the Two-Parameter Beta Method, respectively. In these figures, also presented are the theoretical item characteristic function of item 6, the proportions correct of θ , and the criterion item characteristic function obtained by the Simple Sum Procedure, by solid curves, solid lines with diamonds, and solid curves with crosses, respectively. We can see in these two figures that the results obtained in Degree 3 and 4 Cases are practically identical, in spite of the differences between the two sets of estimated portions of the density function of θ , as we have seen in Figure 5-13-7. This turned out to be

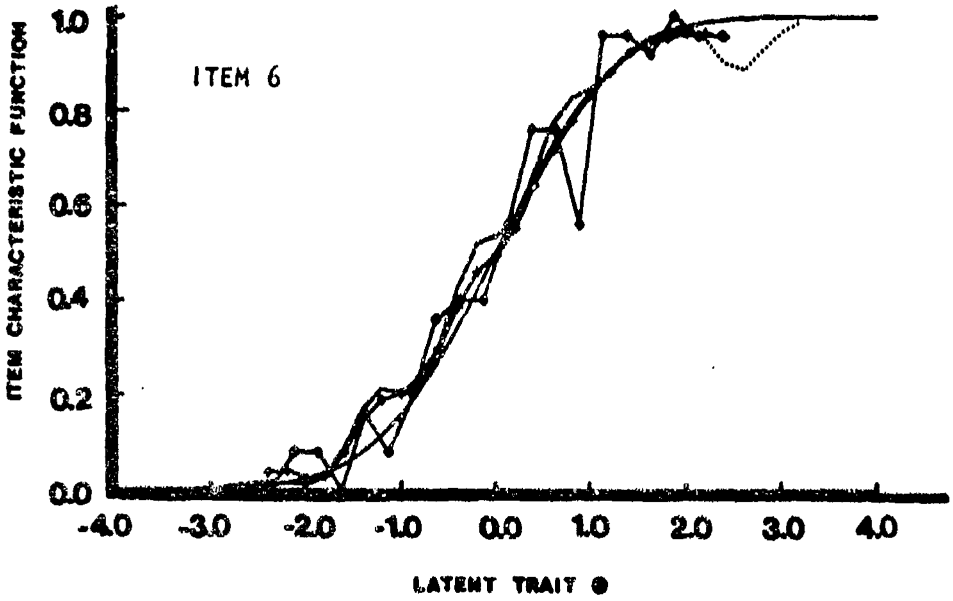


FIGURE 5-13-8

Estimated Item Characteristic Functions of Item 6 in Degree 3 (Dotted Curve) and 4 (Long, Dashed Curve) Cases, Obtained by the Weighted Sum Procedure of the Conditional P.D.F. Approach with the Pearson System Method, in Comparison with the Theoretical Item Characteristic Function (Solid Curve), the Frequency Ratios of θ (Solid Line with Diamonds), and the Criterion Item Characteristic Function in the Simple Sum Procedure (Solid Curve with Crosses).

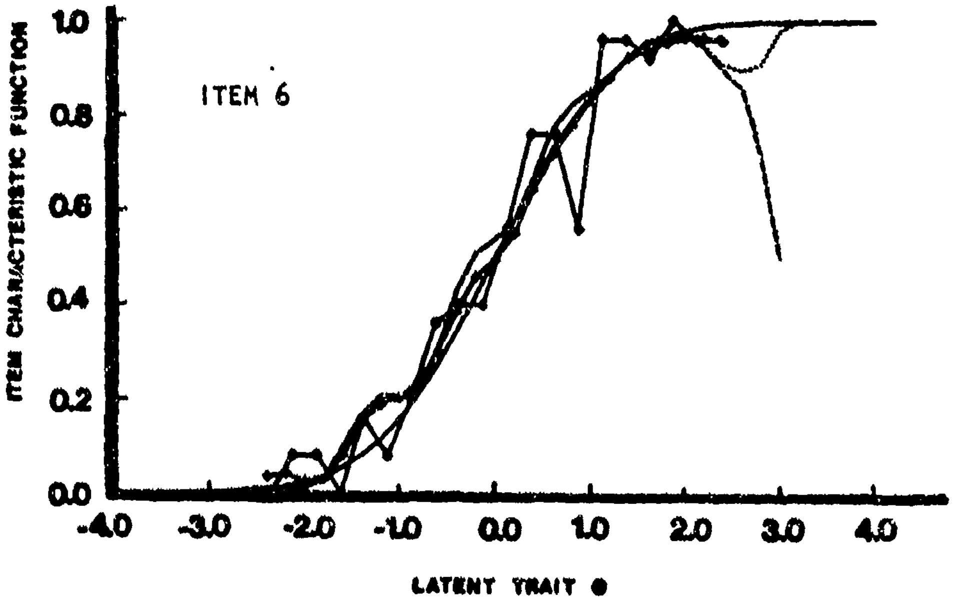


FIGURE 5-13-9

Estimated Item Characteristic Functions of Item 6 in Degree 3 (Dotted Curve) and 4 (Long, Dashed Curve) Cases, Obtained by the Weighted Sum Procedure of the Conditional P.D.F. Approach with the Two-Parameter Beta Method, in Comparison with the Theoretical Item Characteristic Function (Solid Curve), the Frequency Ratios of θ (Solid Line with Diamonds), and the Criterion Item Characteristic Function in the Simple Sum Procedure (Solid Curve with Crosses).

true with every binary test item for the interval of θ , (-2.2, 2.2) , in both results obtained by the Pearson System Method and by the Two-Parameter Beta Method. We also notice that these two sets of results obtained by the two different methods are very close to each other for this range of θ , and, again, this is true with all the other nine binary test items. There are some discrepancies between these results and the criterion item characteristic function obtained by the Simple Sum Procedure, however. Since the estimated item characteristic function obtained by the Simple Sum Procedure with either one of the three methods, i.e., Pearson System Method, Two-Parameter Beta Method and the Normal Approach Method, is practically identical with the corresponding criterion item characteristic functions for each of the ten binary test items, as we have observed earlier in this section, the above discrepancies also exist between the set of estimated item characteristic functions obtained by the Weighted Sum Procedure and the one obtained by the Simple Sum Procedure. In this example of item 6, we can see that the results obtained by the Simple Sum Procedure fit better to the true item characteristic function, than those obtained by the Weighted Sum Procedure. This fact cannot be generalized to all the other nine binary test items, however. For instance, for item 10, the results indicate that this order is reversed.

If we replace $\hat{\phi}(\tau|\hat{\tau}_s)$ in (5.52) by its theoretical counterpart, $\phi(\tau|\hat{\tau}_s)$, which is given by (5.48), we obtain a kind of criterion operating characteristic in the Weighted Sum Procedure. Since we still use the weight obtained from $\hat{g}(\hat{\theta})$ in our example, we shall call it pseudo-criterion item characteristic function. Actually, we can obtain more than one such functions, depending upon the approximations used for $\hat{g}(\hat{\theta})$. We obtained three pseudo-criterion item characteristic functions for each of the ten binary test items, using the three polynomials of degrees 3, 4 and 5, which were obtained by the method of moments and are illustrated in Figure 4-1-2. These three pseudo-criterion item characteristic functions turned out to be very close to the two estimated item characteristic functions of Degree 3 and 4 Cases for each of the ten binary test items, the result which supports the usefulness of the three different methods of approximating the conditional density, $\phi(\tau|\hat{\tau}_s)$.

Proportioned Sum Procedure has a somewhat different rationale from those for the other two procedures. Let $p(s\varepsilon x_h)$ be the probability with which the examinee s belongs to the subpopulation x_h . We have for the estimated operating characteristics, $\hat{P}_{x_h}(\theta)$, of the item response x_h to item h .

$$(5.53) \qquad \hat{P}_{x_h}(\theta) = \sum_{s=1}^{N} \hat{p}(s\varepsilon x_h)\, \hat{\phi}(\tau|\hat{\tau}_s)\, [\sum_{s=1}^{N} \hat{\phi}(\tau|\hat{\tau}_s)]^{-1} \ ,$$

$$x_h = 0,1,\dots,m_h \ ,$$

where $\hat{p}(s\epsilon x_h)$ is the estimate of the probability $p(s\epsilon x_h)$, which satisfies

$$(5.54) \qquad \sum_{x_h=0}^{m_h} \hat{p}(s\epsilon x_h) = \sum_{x_h=0}^{m_h} p(s\epsilon x_h) = 1 \quad .$$

Figure 5-13-10 presents the four different estimates of $p(s\epsilon x_h)$ for item 6, which were used in the present study. Our basic data are, again, the set of five hundred maximum likelihood estimates obtained upon the original Old Test. These four estimates of $p(s\epsilon x_h)$ are the proportions of the examinees who belong to the subpopulation x_h within a more or less arbitrarily chosen interval of $\hat{\theta}$. The first and second estimates, which are plotted by solid triangles and crosses, respectively, in

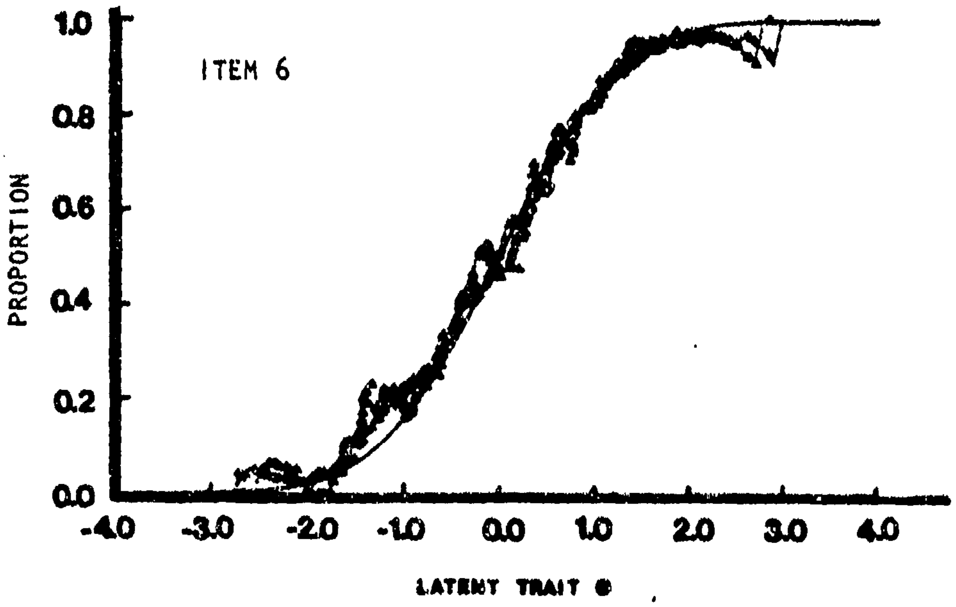


FIGURE 5-13-10

Four Different Estimates of $p(s\epsilon x_h=1)$ for Item 6, i.e., the Proportions of the Examinees Who Answered Correctly to Item 6 within the Interval $\hat{\theta}_s \pm \sigma$ (Solid Triangles), Those within the Interval $\hat{\theta}_s \pm 2\sigma$ (Crosses), and the Corresponding Results for Which the 61 Equally Spaced Values of θ Were Used Instead of the 500 Values of $\hat{\theta}_s$ (Dots and Dashes, Respectively).

Figure 5-13-10, are the proportions of the examinees who belong to the subpopulation $x_h=1$ within the intervals, $\hat{\theta}_s \pm \sigma$ and $\hat{\theta}_s \pm 2\sigma$, respectively, where $\sigma = 0.215$. The third and fourth ones, which are drawn by dots and dashes, respectively, are the same as the first two, but are assigned to the sixty-one equally spaced values of $\hat{\theta}$, instead of the five hundred observations, $\hat{\theta}_s$'s . We notice that these proportions themselves can be crude estimates of the operating characteristic of x_h, if we correct the scale of $\hat{\theta}$ using the method suggested in Section V.2 . With our data, the ratio of the standard deviation of $\hat{\theta}$ to that of θ is only 1.011 (cf. RR-78-5) and the regression of θ on $\hat{\theta}$ is approximately linear for the interval of $\hat{\theta}$, (-2.2, 2.0) (cf. Figure 5-13-4). For these reasons, the item characteristic function of item 6 is drawn without correction in Figure 5-13-10, for a rough comparison.

Figures 5-13-11 and 5-13-12 present the resultant estimated item characteristic functions of item 6 obtained by the Proportioned Sum Procedure which is combined with the Pearson System Method and the Two-Parameter Beta Method, respectively, using the first two $\hat{p}(s\varepsilon x_h)$'s , for Degree 3 and 4 Cases. In these figures, the results obtained by using the first and second $\hat{p}(s\varepsilon x_h=1)$'s for Degree 3 Case are plotted by dots and medium dashes, and those for Degree 4 Case are drawn by short and long dashes, respectively, together with the theoretical item characteristic function of item 6, the proportions correct of θ, and the criterion item characteristic function obtained by the Simple Sum Procedure, which are drawn by solid curves, lines with diamonds, and curves with crosses, respectively. We can see in each of these two figures that the four results are very close to each other, and also to the criterion item characteristic function obtained by the Simple Sum Procedure, for the interval of θ, (-2.5, 2.5) . This is a common tendency among all the ten binary test items, although for some items they are not as close as those for item 6. It is also noted that these two sets of results obtained by the Pearson System Method and by the Two-Parameter Beta Method are very close to each

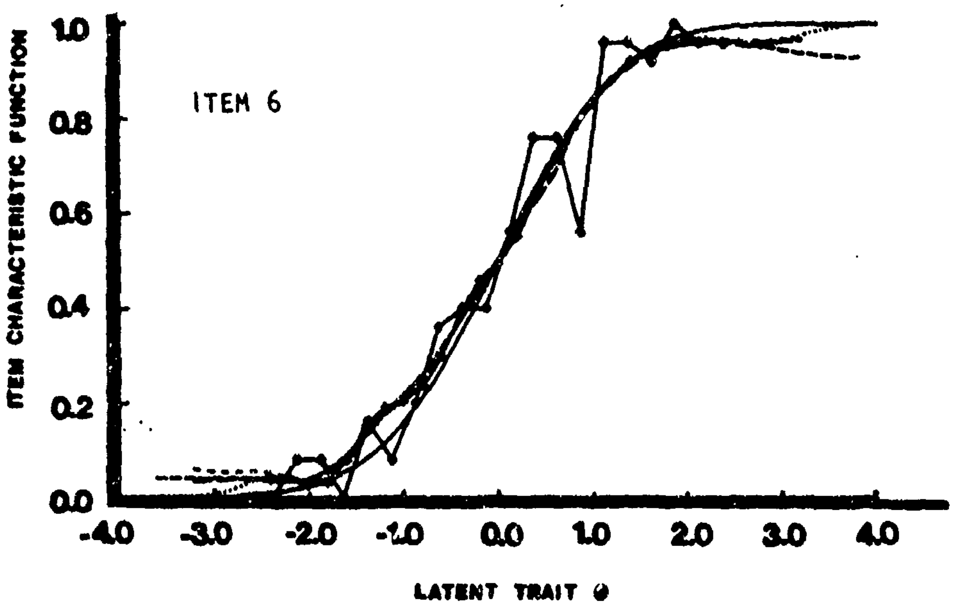


FIGURE 5-13-11

Estimated Item Characteristic Functions of Item 6 Obtained by the Proportioned Sum Procedure of the Conditional P.D.F. Approach with the Pearson System Method, by Using the Proportions for $\hat{\theta}_s \pm \sigma$ in Degree 3 (Dots) and 4 (Short Dashes) Cases, and by Using Those for $\hat{\theta}_s \pm 2\sigma$ in Degree 3 (Medium Dashes) and 4 (Long Dashes) Cases, Respectively. They Are Compared with the Theoretical Item Characteristic Function (Solid Curve), the Frequency Ratios of θ (Solid Line with Diamonds), and the Criterion Item Characteristic Function in the Simple Sum Procedure (Solid Curve with Crosses).

other. This tendency is common to all the ten binary test items.

If we replace $\hat{\phi}(\tau|\hat{\tau}_s)$ in (5.53) by the true density, $\phi(\tau|\hat{\tau}_s)$, we can obtain the pseudo-criterion operating characteristic of x_h. In the present study, four different pseudo-criterion item characteristic functions were obtained, using the four different estimates of $p(s\varepsilon x_h=1)$, which we have observed in Figure 5-13-10. The resultant pseudo-criterion item characteristic functions turned out to be very close to the estimated item characteristic functions obtained by using the same $\hat{p}(s\varepsilon x_h=1)$, for each of the ten binary test items, the fact which supports the usefulness of both Pearson System Method and Two-Parameter Beta Method.

The estimated ability distributions for the success and the

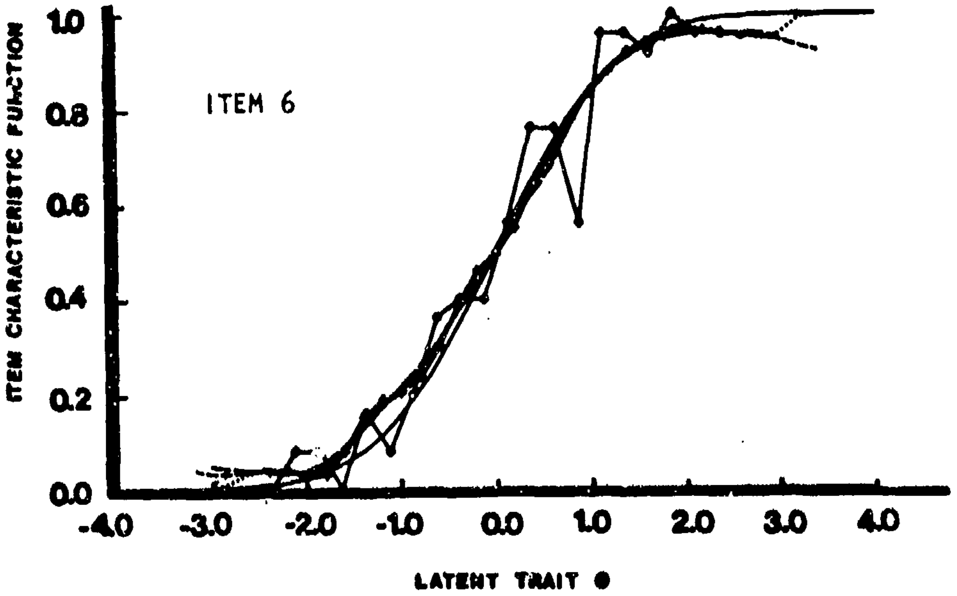


FIGURE 5-13-12

Estimated Item Characteristic Functions of Item 6 Obtained by the Proportioned Sum Procedure of the Conditional P.D.F. Approach with the Two-Parameter Beta Method, by Using the Proportions for $\hat{\theta}_s \pm \sigma$ in Degree 3 (Dots) and 4 (Short Dashes) Cases, and by Using Those for $\hat{\theta}_s \pm 2\sigma$ in Degree 3 (Medium Dashes) and 4 (Long Dashes) Cases, Respectively. They Are Compared with the Theoretical Item Characteristic Function (Solid Curve), The Frequency Ratios of θ (Solid Line with Diamonds), and the Criterion Item Characteristic Function in the Simple Sum Procedure (Solid Curve with Crosses).

failure subpopulations for each item turned out to be very similar to those obtained by the other combinations of an approach and a method, for both Degree 3 and 4 Cases.

Figure 5-13-13 presents the estimated density functions for the total population, which were obtained by the Two-Parameter Beta Method, using $\hat{\theta}_s \pm 0.215$ as the interval for computing $\hat{p}(s\varepsilon x_h=1)$, in Degree 3 and 4 cases, by dotted and dashed curves, respectively, together with the theoretical density, $f(\theta)$. We can see in this figure that these two results are close to each other, and reasonably close to the uniform density. The corresponding results obtained by using the interval, $\hat{\theta}_s \pm 0.430$, turned out to be very close to these results.

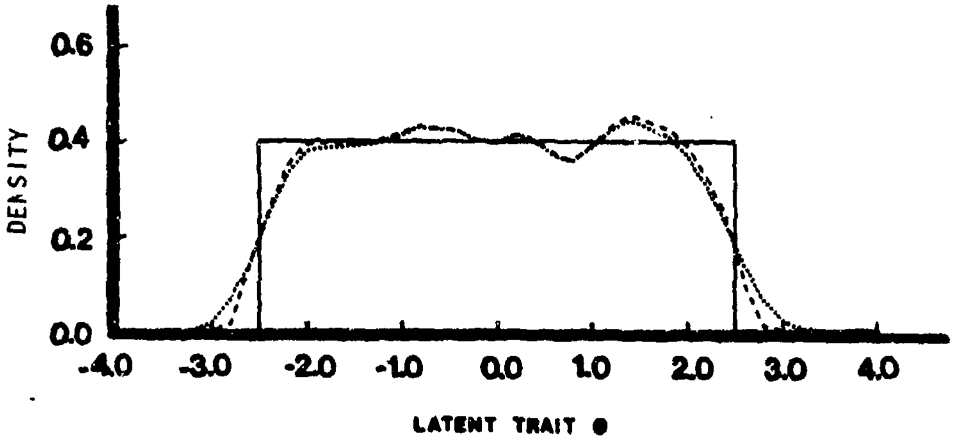


FIGURE 5-13-13

Estimated Density Functions of Ability θ , Obtained by the Proportioned Sum Procedure of the Conditional P.D.F. Approach with the Two-Parameter Beta Method, by Using the Proportions for the Interval, $\hat{\theta}_s \pm \sigma$, in Degree 3 (Dots) and Degree 4 (Short Dashes) Cases, in Comparison with the Theoretical Density Function.

We have also obtained the corresponding four estimated density functions by the Pearson System Method. The results turned out to be fairly close to those obtained by the Two-Paramcter Beta Method. In fact, all the other results obtained by the oth approaches turned out to be similar, with some deviations, i.e., some of them are a little closer to the theoretical density function, and some of them are a little less close.

(V.14) Remark on the Approximation of $\phi(\tau|\hat{\tau})$ by a Normal Density Function

We have seen in the previous sections that, in spite of its relatively restricted shape of the normal density dunction, -mal Approach Method works just as well as the other two methods, i.e., Pearson System Method and Two-Parameter Beta Method, in approximating the conditional density function, $\phi(\tau|\hat{\tau})$. There is a good reason behind this fact, which we shall observe in this section.

Suppose that the density function, $f^*(\tau)$, is uniform for a certain interval of τ , $[\underline{\tau},\bar{\tau}]$. Then we can write

$$(5.55) \qquad \phi(\tau|\hat{\tau}) = \psi(\hat{\tau}|\tau)\ f^*(\tau)\ [\int_{\underline{\tau}}^{\bar{\tau}} \psi(\hat{\tau}|\tau)\ f^*(\tau)\ d\tau]^{-1}$$

$$= \psi(\hat{\tau}|\tau)\ [\int_{\underline{\tau}}^{\bar{\tau}} \psi(\hat{\tau}|\tau)\ d\tau]^{-1} \quad , \text{ for } \underline{\tau} < \tau < \bar{\tau} \ .$$

Since we have

$$(5.56) \qquad \psi(\hat{\tau}|\tau) = (2\pi)^{-1/2}\ \sigma^{-1}\ \exp[(\hat{\tau}|\tau)^2/(2\sigma^2)]$$

$$= (2\pi)^{-1/2}\ \sigma^{-1}\ \exp[(\tau|\hat{\tau})^2/(2\sigma^2)] \ ,$$

from this and (5.55), we find that $\phi(\tau|\hat{\tau})$ is a truncated normal density function. When σ is small, for a wide range of $\hat{\tau}$, this is practically equal to the complete normal density function, which is given by the rightest-hand side of (5.56). Normal Approach Method, therefore, must work well in this situation.

If the marginal density, $f^*(\tau)$, is a normal density function with μ and ζ as its two parameters, then the joint distribution of τ and $\hat{\tau}$ will be the bivariate normal distribution, with μ and $(\sigma^2+\zeta^2)^{1/2}$ as the two parameters for the marginal density function, $g(\hat{\tau})$, and

$$(5.57) \qquad \rho = \zeta(\sigma^2+\zeta^2)^{-1/2}$$

as the fifth parameter. Thus the conditional density, $\phi(\tau|\hat{\tau})$, is a normal density function, with $(\zeta^2\hat{\tau}+\sigma^2\mu)(\sigma^2+\zeta^2)^{-1}$ and $\sigma\zeta(\sigma^2+\zeta^2)^{-1/2}$ as the two parameters.

These two facts indicate that, if the distribution of τ is close to either a normal distribution or a uniform distribution, or between the two, Normal Approach Method will work well in approximating the conditional density function, $\phi(\tau|\hat{\tau})$.

REFERENCES

[1] Elderton, W. P. and N. L. Johnson. Systems of frequency curves. Cambridge University Press, 1969.

[2] Johnson, N. L. and S. Kotz. Continuous univariate distributions. Vol. 2. Houghton Mifflin, 1970.

[3] Samejima, F. A method of estimating item characteristic functions using the maximum likelihood estimate of ability. Psychometrika, 42, 1977, pages 163-191.

VI Estimation of the Operating Characteristics of the Discrete Item Responses and That of Ability Distributions: III

Following Chapters 3 and 5, in the present chapter, we shall integrate the results and findings of the part of our research under this title. It also includes certain observations about tests in general, objective testing, ethics behind Bayesian estimation, and some other related topics. The main subject in the present chapter is to find out how small the number of test items can be in our Old Test. Alternative estimators for the maximum likelihood estimator will be introduced, which can be used when the amount of test information of our Old Test is not large enough for the entire range of ability θ of our interest, and, consequently, there exist more than a few positive and/or negative infinities for the maximum likelihood estimates of ability of our examinees.

(VI.1) Objective Testing and Exchangeabil

Equal opportunities have been considered to be ethical in our society. In personnel selection, for example, we are supposed to make our decisions which are based upon the applicants' capabilities, but not upon their ethnic backgrounds, ages, sexes, and other attributes which have little to do with their capabilities for a specified job. The translation of this equal opportunity principle to testing will be that we should: 1) develop and use valid tests for the selection purpose; 2) objectively analyze the results of the tests; and 3) make our recommendations as to which applicants should be accepted and which should be rejected on the basis of these objective findings only.

Although the above first and third statements are readily accepted by people in general, including researchers, for some reason the second statement has attracted little attention. Note, however, that this is the part that researchers should be most responsible for.

Bayesian estimation of ability has been accepted for many years as a valid method by researchers. This fact does not justify, however, certain serious flaws Bayesian estimation has, which are clearly against the principle of objective testing. It assumes the exchangeability of individuals who belong to a certain subpopulation, and uses the ability distribution of the subpopulation as the prior.

Let us assume that we have two ethnic groups, A and B . Figure 6-1-1 presents the priors of these two hypothetical ethnic subpopulations with respect to ability θ . The basic idea behind the Bayesian estimation is that, within each ethnic subpopulation, A or B , the individuals are exchangeable. Are they really? Suppose that we fix the level of θ at θ_0 , as is indicated in Figure 6-1-1. If we consider the subset of individuals whose ability levels are uniformly θ_0 , they will include certain people from the ethnic group A , and also certain other people who belong to B . Our best common sense tells us that these individuals are the people who are exchangeable. In the Bayesian estimation, however, they are not; in its logic, those who belong to A are exchangeable among themselves, and so are those who belong to B .

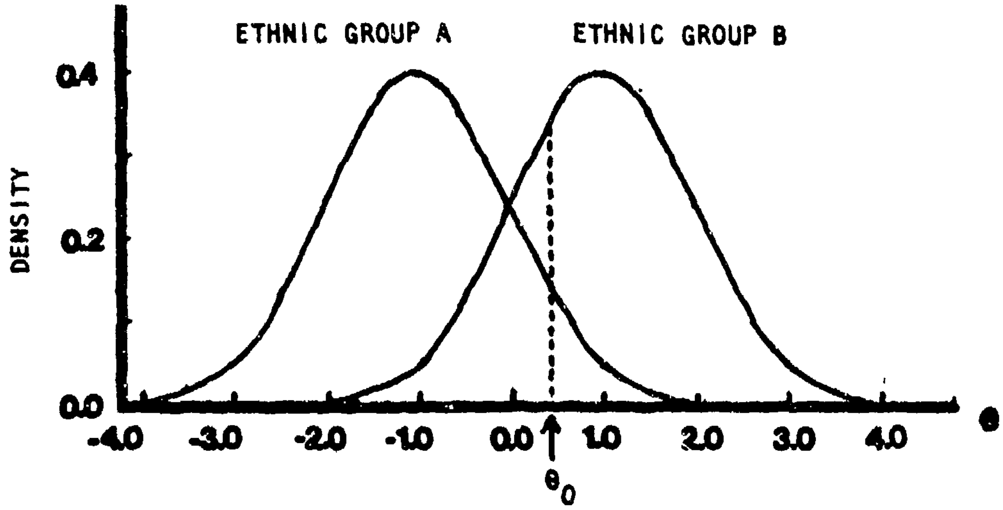


FIGURE 6-1-1

Density Functions of the Ability Distributions of Two Hypothetical Ethnic Groups, A and B .

In order to observe this issue from a somewhat different angle, we shall consider a real test, LIS-U (Indow and Samejima, 1962, 1966; Samejima, 1969, RR-80-3). Figure 6-1-2 presents the test information function and its square root of LIS-U by solid and

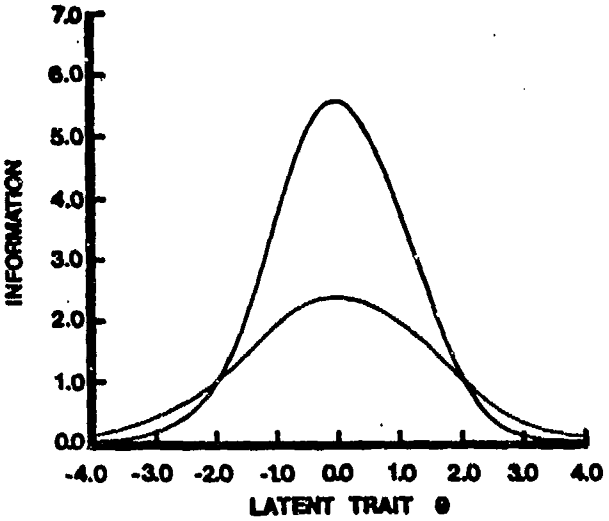


FIGURE 6-1-2

Test Information Function (Solid Line) and Its Square Root (Dotted Line) of LIS-U.

dotted curves, respectively. The test consists of seven binary items, which make a fairly short test. Bayes modal estimator (Samejima, 1969), $\hat{\theta}_V$, of ability θ is the modal point of θ for the function $B_V(\theta)$ such that

$$(6.1) \qquad B_V(\theta) = P_V(\theta)\ f(\theta) \quad .$$

This estimator was adopted as the estimator of ability θ , using the density function of the ability distribution, $f(\theta)$, as the prior, for each of the $2^7 = 128$ response patterns. The regression of $\hat{\theta}$ on ability θ is given by

$$(6.2) \qquad E[\hat{\theta}|\theta] = \sum_V \hat{\theta}_V\ P_V(\theta) \quad .$$

Figure 6-1-3 presents the four regressions of $\hat{\theta}$ on θ using the four different priors, n(0.0,1.0), n(-1.0,1.0), n(1.0,1.0) and n(0.0,0.5) , by solid, long dashed, short dashed and dotted curves, respectively. Let us assume that the second prior is for the ethnic group A , and the third prior is for the ethnic group B , and $\theta_0 = 2.0$. We can see in this figure that, for two individuals, whose ability levels are uniformly θ_0 but belong to the ethnic groups A and B , respectively, the distributions of the Bayes modal estimate, $\hat{\theta}$, are different, and their expectations are approximately 1.0 and 1.6 , respectively --- a substantial difference!

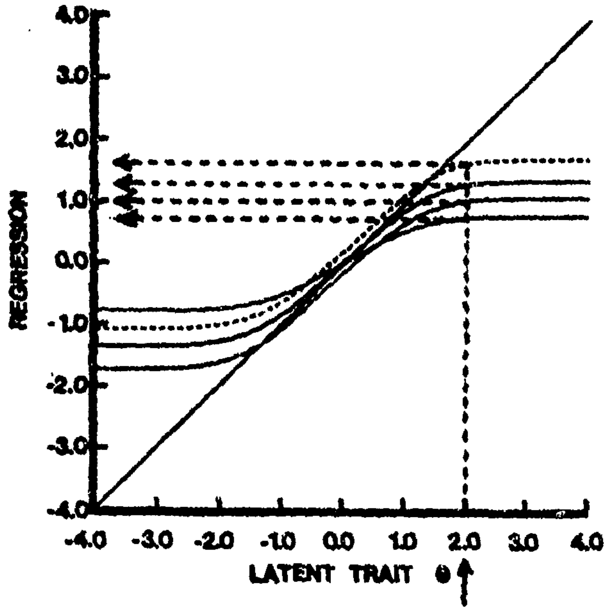


FIGURE 6-1-3

Four Regressions of the Bayes Modal Estimate on Ability Based on LIS-U , with the Priors, n(0.0,1.0) (Solid Line) , n(-1.0,1.0) (Long, Dashed Line), n(1.0,1.0) (Short, Dashed Line) , and n(0.0,0.5) (Dotted Line) , Respectively.

Let us assume, further, that the first and the fourth priors are for men and women, respectively. If the first individual of the above two happens to be male, then, using n(0.0,1.0) as the prior, his expected Bayes modal estimate is approximately 1.3 . Which should we take for this first individual, 1.0 or 1.3 , as the expected value of his ability estimate? This individual will obtain a higher score if the prior for men is used than if that of the ethnic group A is used. Perhaps he would rather be treated as a man than as a member of the ethnic group A . If the second individual happens to be female, then her expected Bayes modal estimate becomes approximately 0.7 . Again, there are two expected values for her, 1.6 and 0.7 , and she will obtain a higher score if she is categorized as a member of the ethnic group B rather than as a woman. If we use the second priors for the two individuals, the expected Bayes modal estimates are 1.3 and 0.7 , i.e., the reversal of the order from that of 1.0 and 1.6 ! Thus, if we take the first set of priors in selection, then we will be saying, "If there are two people whose ability levels are exactly the same and at 2.0 , then we will accept the one from the ethnic group B ." If we take the second set of priors, then we will be saying, "If there are two such individuals who happen to be male and female, then we will accept the man and reject the woman." We will be very likely to accept the second individual if we take the first set of priors, and, if we take the second set, then it will be highly probable that we accept the first individual and reject the second. This is what it amounts to when we use a Bayesian estimator of ability in our selection.

A solution for this chaos will be to divide each ethnic group further, to make four groups instead of two, i.e., ethnic A and male, ethnic B and male, ethnic A and female, and ethnic B and female. It should be noted, however, that every individual has much more than two casual attributes like his or her ethnic background and sex, and similar problems will happen for these four groups. Then we may need eight groups instead of four, sixteen, thirty-two, etc. In this way, we will reach, fairly soon, the conclusion that each individual has his or her own prior, or each prior includes only one individual. Then Bayesian estimation may finally be justifiable and useful. In such a case, however, why do we need testing at all if we know about each individual's ability so well? In most cases we do not, and that is why we need testing.

The flaw of the Bayesian estimation comes from the fact that it deals with a group of individuals who are not exchangeable as if they were exchangeable, and treats those who are exchangeable, i.e., individuals whose ability levels are exactly the same, as if they were not exchangeable. This is against the principle of objective testing. It is a typical example of failure in objectively analyzing the results of testing.

(VI.2) Every Test Has a Limitation

We can see in Figure 6-1-3 that for the values of ability θ, approximately, greater than 1.0 and also those, approximately, less than -1.0, there are little changes in the regression of $\hat{\theta}$ on θ, for each of the four different priors. In fact, the conditional distribution of the Bayes modal estimate, $\hat{\theta}$, given θ, approaches the one point distribution at the modal point of θ for the product, $P_{V\text{-min}}(\theta)\ f(\theta)$, as θ tends to negative infinity, and it approaches the one point distribution at the modal point of θ for the product, $P_{V\text{-max}}(\theta)\ f(\theta)$, as θ tends to positive infinity, where V-min and V-max indicate the two extreme response patterns, $(0,0,\ldots,0)$ and $(m_1,m_2,\ldots,m_n)$. This means that for these outside ranges of ability θ LIS-U is powerless, and it is the prior that takes the essential role in determining the value of the Bayes modal estimate. It is as if the examinee were cheated, obtaining something other than the information the test itself has provided.

We must accept the fact that every test has a limitation as to the range of ability which it can measure. Escaping to priors will by no means enhance this range, but will impose the bias which was described in the preceding section. No single test has an infinite number of test items, so it should not be expected that any test can measure an unlimited range of ability.

(VI.3) Alternative Estimators for the Maximum Likelihood Estimator

A question will arise as to whether there is any way to enhance the range of ability for which a specified test is powerful, without depending upon priors or any other resources of irrelevant information. This can be done by replacing negative and positive infinities of the maximum likelihood estimates for the two extreme response patterns, V-min and V-max , by some appropriate finite numbers. In search of such alternative estimators, our goal was to find suitable substitutes which do not depend upon any specific populations of examinees, but are population-free, unlike Bayesian estimators.

It is desirable that such alternative estimators provide us with the conditional unbiasedness, given θ , as is the case with the maximum likelihood estimator in the limiting situation where we have infinitely many test items. We notice that the operating characteristic $P_{V\text{-min}}(\theta)$ strictly decreases in θ , and $P_{V\text{-max}}(\theta)$ strictly increases in θ , as long as our test items follow a model, or models, like the normal ogive and logistic models. Thus we can conceive of a critical point, θ_c , which satisfies

$$(6.3) \qquad \begin{cases} P_{V\text{-min}}(\theta) \doteq 0 & \text{for } \theta > \theta_c \\ P_{V\text{-max}}(\theta) \doteq 0 & \text{for } \theta \le \theta_c \end{cases} .$$

Figure 6-3-1 presents the operating characteristics of the two extreme response patterns, $P_{V\text{-min}}(\theta)$ and $P_{V\text{-max}}(\theta)$, of LIS-U , by solid and dotted curves, respectively. The critical value, θ_c , was obtained in such a way that the product of these two operating characteristics be maximal at this point. It turned out to be -0.0088 .

We shall aim at finding finite substitutes for the two maximum likelihood estimates, $\hat{\theta}_{V\text{-min}}$ and $\hat{\theta}_{V\text{-max}}$, which are negative and positive infinites, respectively, in such a way that the substitution should provide us with a regression which is close

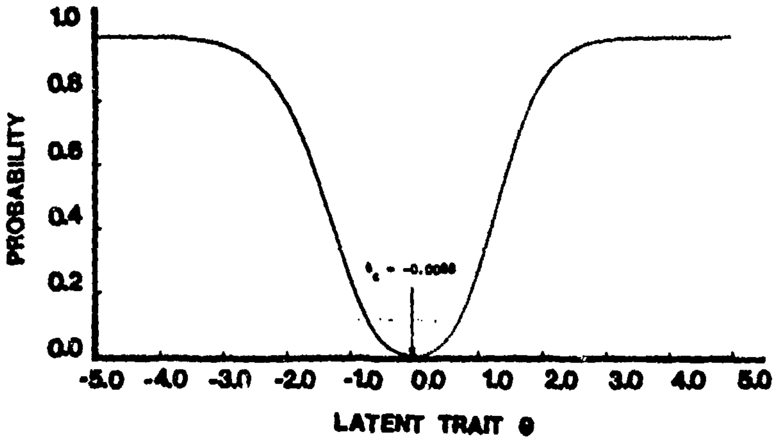


FIGURE 6-3-1

Operating Characteristics of the Two Extreme Response Patterns, (0,0,0,0,0,0,0) (Solid Line) and (1,1,1,1,1,1,1) (Dotted Line), of LIS-U , and the Position of the Critical Value θ_c .

enough to θ , i.e., the unbiasedness of the estimator, for some range of θ . Let $\theta^*_{V\text{-min}}$ and $\theta^*_{V\text{-max}}$ denote such estimates, and θ^*_V be the resultant estimator, such that

$$(6.4) \qquad \theta^*_V \begin{cases} = \theta^*_{V\text{-min}} & \text{for } V\text{-min} \\ = \theta^*_{V\text{-max}} & \text{for } V\text{-max} \\ = \hat{\theta}_V & \text{for all the other response patterns.} \end{cases}$$

We can write for the regression of θ^*_V on ability θ such that

$$(6.5) \qquad E(\theta^*_V|\theta) = \sum_{\substack{V \neq V\text{-min} \\ V \neq V\text{-max}}} \hat{\theta}_V \, P_V(\theta) + \theta^*_{V\text{-min}} \, P_{V\text{-min}}(\theta)$$

$$+ \theta^*_{V\text{-max}} \, P_{V\text{-max}}(\theta)$$

$$\begin{cases} \doteq \displaystyle\sum_{\substack{V\neq V\text{-min}\\ V\neq V\text{-max}}} \hat{\theta}_V\, P_V(\theta) + \theta^*_{V\text{-min}}\, P_{V\text{-min}}(\theta) & \\ \qquad\qquad\qquad \text{for} \quad \theta \leq \theta_c & \\ \doteq \displaystyle\sum_{\substack{V\neq V\text{-min}\\ V\neq V\text{-max}}} \hat{\theta}_V\, P_V(\theta) + \theta^*_{V\text{-max}}\, P_{V\text{-max}}(\theta) & \\ \qquad\qquad\qquad \text{for} \quad \theta > \theta_c \ . & \end{cases}$$

If this estimator, θ^*_V , provides us with an approximate unbiasedness for a certain range of θ , $(\underline{\theta},\bar{\theta})$, then we shall be able to write

$$(6.6) \quad \begin{cases} \displaystyle\sum_{\substack{V\neq V\text{-min}\\ V\neq V\text{-max}}} \hat{\theta}_V\, P_V(\theta) + \theta^*_{V\text{-min}}\, P_{V\text{-min}}(\theta) \doteq \theta & \\ \qquad\qquad\qquad \text{for} \quad \theta < \theta \leq \theta_c & \\ \displaystyle\sum_{\substack{V\neq V\text{-min}\\ V\neq V\text{-max}}} \hat{\theta}_V\, P_V(\theta) + \theta^*_{V\text{-max}}\, P_{V\text{-max}}(\theta) \doteq \theta & \\ \qquad\qquad\qquad \text{for} \quad \theta_c < \theta < \bar{\theta} \ . & \end{cases}$$

In practice, we must search the interval of θ , $(\underline{\theta},\bar{\theta})$, for which such an estimator, θ^*_V , is available, in relation with a specific test of our interest. From (6.6), we can further write

$$(6.7) \quad \begin{cases} \displaystyle\sum_{\substack{V\neq V\text{-min}\\ V\neq V\text{-max}}} \hat{\theta}_V \int_{\underline{\theta}}^{\theta_c} P_V(\theta)\, d\theta + \theta^*_{V\text{-min}} \int_{\underline{\theta}}^{\theta_c} P_{V\text{-min}}(\theta)\, d\theta & \\ \qquad \doteq \frac{1}{2}\,(\theta_c^2 - \bar{\theta}^2) & \\ \displaystyle\sum_{\substack{V=V\text{-min}\\ V=V\text{-max}}} \hat{\theta}_V \int_{\theta}^{\bar{\theta}_c} P_V(\theta)\, d\theta + \theta^*_{V\text{-max}} \int_{\theta_c}^{\bar{\theta}_c} P_{V\text{-max}}(\theta)\, d\theta & \\ \qquad \doteq \frac{1}{2}\,(\bar{\theta}^2 - \theta_c^2) \ . & \end{cases}$$

Thus the two estimates, θ^*_{V-min} and θ^*_{V-max} , can be obtained by

$$(6.8)\quad \left\{ \begin{aligned} \theta^*_{V-min} &= [\tfrac{1}{2}(\theta_c^2 - \underline{\theta}^2) - \sum_{\substack{V\neq V-min \\ V\neq V-max}} \hat{\theta}_V \int_{\underline{\theta}}^{\theta_c} P_V(\theta)\, d\theta] \\ &\qquad [\int_{\underline{\theta}}^{\theta_c} P_{V-min}(\theta)\, d\theta]^{-1} \\ \theta^*_{V-max} &= [\tfrac{1}{2}(\bar{\theta}^2 - \theta_c^2) - \sum_{\substack{V\neq V-min \\ V\neq V-max}} \hat{\theta}_V \int_{\theta_c}^{\bar{\theta}} P_V(\theta)\, d\theta] \\ &\qquad [\int_{\theta_c}^{\bar{\theta}} P_{V-max}(\theta)\, d\theta]^{-1} \end{aligned} \right. ,$$

with some appropriate values for $\underline{\theta}$ and $\bar{\theta}$.

We used eleven different sets of $\underline{\theta}$ and $\bar{\theta}$, ±1.50 , ±1.75 , ±2.00 , ±2.25 , ±2.50 , ±3.00 , ±3.50 , ±4.00 , ±4.50 , ±5.00 , and ±5.50 , for the purpose of experimentation. The resultant set of estimates, θ^*_{V-min} and θ^*_{V-max} , which was obtained by using each of these eleven intervals, is given in Table 6-3-1. Figure 6-3-2 illustrates the regressions of θ^*_{V-min} on θ , obtained by using (-1.5, 1.5) and (-2.25, 2.25) , respectively, as $(\underline{\theta},\bar{\theta})$, by solid and dashed curves. The values of θ^*_{V-min} and θ^*_{V-max} turned out to be -1.47883 and 1.52237 in the former case, and -1.79255 and 1.77649 in the latter, as we can see in Table 6-3-1. In the same figure also presented are the unbiasedness line, i.e., the line which passes the origin with the angle of 45 degree from the abscissa, and the regression of the Bayes modal estimate with the prior, n(0.0,1.0) , by a solid line and a dotted curve, respectively. We can see in this figure that, within each interval, each of these two regressions is reasonably close to the unbiasedness line, and much closer than the regression of the Bayes modal estimate. If we enhance the interval further, the deviation

TABLE 6-3-1

Eleven Sets of Estimates, $\theta^*_{V\text{-min}}$ and $\theta^*_{V\text{-max}}$, of Ability for the Two Extreme Response Patterns, (0,0,...,0) and (1,1,...,1), Obtained on LIS-U, Using Eleven Different Intervals for $(\underline{\theta},\overline{\theta})$.

$\underline{\theta}$, $\overline{\theta}$	$\theta^*_{V\text{-min}}$	$\theta^*_{V\text{-max}}$
± 1.50	-1.47883	1.52237
± 1.75	-1.64702	1.65605
± 2.00	-1.79255	1.77649
± 2.25	-1.92540	1.89233
± 2.50	-2.05136	2.00754
± 3.00	-2.29490	2.24127
± 3.50	-2.53641	2.48011
± 4.00	-2.77945	2.72254
± 4.50	-3.02430	2.96720
± 5.00	-3.27051	3.21329
± 5.50	-3.51765	3.46032

from the unbiasedness line becomes larger (cf. RR-80-3). Since the least finite value of the maximum likelihood estimate for LIS-U is -1.3167 for the response pattern, (0,0,0,1,0,0,0) , and the greatest finite value is 1.3028 for (1,1,1,0,1,1,1) , either one of the above sets of $\theta^*_{V\text{-min}}$ and $\theta^*_{V\text{-max}}$ will be adequate, and so is any of them obtained by using intervals between (-1.5, 1.5) and (-2.25, 2.25) .

The introduction of the new estimator, θ^*_V , has enhanced the range of ability for which a given test is meaningful without sacrificing the objectivity of testing, as Bayesian estimates do. When the number of items is as small as seven and all items are binary items, as is the case with LIS-U , the computation of $\theta^*_{V\text{-min}}$ and $\theta^*_{V\text{-max}}$ is relatively easy, owing to the fact that the number of all possible response patterns is as small as 128 .

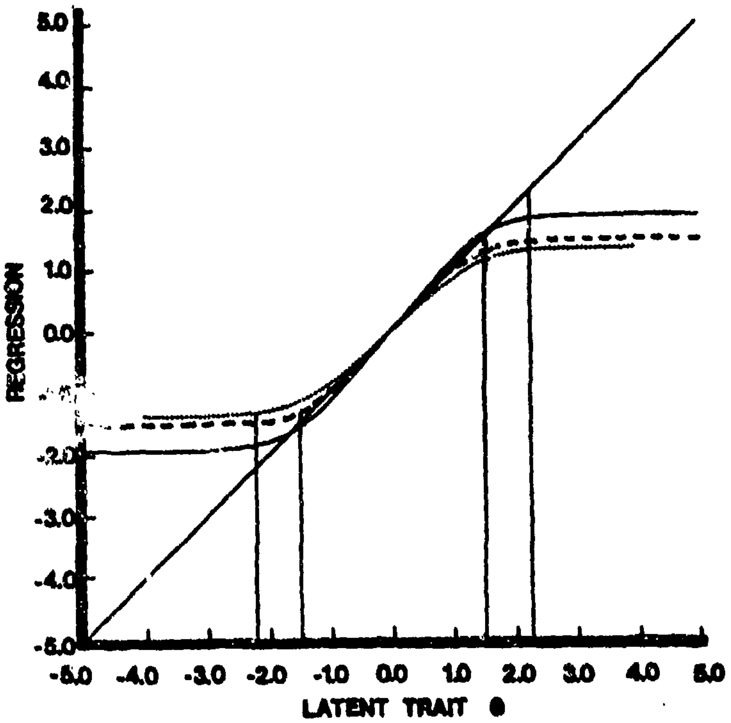


FIGURE 6-3-2

Two Regressions of the Modified Maximum Likelihood Estimate, $\hat{\theta}_V$, on Ability θ , Using (-1.5, 1.5) (Dashed Curve) and (-2.25, 2.25) (Solid Curve) as $(\underline{\theta}, \bar{\theta})$, Together with the Regression of the Bayes Modal Estimate with n(0,1) as the Prior (Dotted Curve) .

Note, however, that the increase in the number of items, and/or in the number of item scores for each item, will soon make it practically impossible to compute these two substitutes estimates, since the number of all possible response patterns will increase by gigantic steps. For example, if a test has ten binary items instead of seven, the number of all possible response patterns will be 1,024 ; if a test has seven three-item-score-category items, the number of all possible response patterns will be 2,187 ; if a test has fifteen three-item-score-category items, it will be as large as 14,348,907 !

It is necessary, therefore, that we invent some method to deal with the situation in which the number of all the possible response patterns is too large for us to compute $\theta^*_{V\text{-min}}$ and $\theta^*_{V\text{-max}}$ directly. By virtue of the availability of electronic computers and the Monte Carlo method, this can be done by introducing the sample statistic versions of the two estimators.

Let N be the number of examinees who were selected randomly from the uniform distribution for the interval of θ, $(\underline{\theta}, \bar{\theta})$. Let N_L denote the number of examinees who belong to the above sample and whose levels of ability are lower than the critical value θ_c, and N_H be of that of those whose ability levels are higher than, or equal to, θ_c. Thus we can write

$$(6.9) \qquad N = N_L + N_H \ .$$

Let N_{LV} and N_{HV} denote the numbers of examinees who obtained the response pattern V, in the above two subgroups of the sample, respectively. Thus we have

$$(6.10) \qquad \begin{cases} N_L = \sum\limits_V N_{LV} \\ N_H = \sum\limits_V N_{HV} \ . \end{cases}$$

It can be seen that the sample statistic corresponding to $\int_{\underline{\theta}}^{\theta_c} P_V(\theta)\, d\theta$ in the formula (6.8) is $N_{LV}(\theta_c - \underline{\theta})\, N_L^{-1}$, and also the one for $\int_{\theta_c}^{\bar{\theta}} P_V(\theta)\, d\theta$ is $N_{HV}(\bar{\theta} - \theta_c)\, N_H^{-1}$. Substituting these sample statistics into (6.8) and rearranging, we obtain $\hat{\theta}^*_{V\text{-min}}$ and $\hat{\theta}^*_{V\text{-max}}$ such that

$$(6.11) \qquad \begin{cases} \hat{\theta}^*_{V\text{-min}} = [\frac{1}{2}(\theta_c + \underline{\theta})\, N_L - \sum\limits_{\substack{V \neq V\text{-min} \\ V \neq V\text{-max}}} \hat{\theta}_V\, N_{LV}]\, N^{-1}_{LV\text{-min}} \\ \hat{\theta}^*_{V\text{-max}} = [\frac{1}{2}(\bar{\theta} + \theta_c)\, N_H - \sum\limits_{\substack{V \neq V\text{-min} \\ V \neq V\text{-max}}} \hat{\theta}_V\, N_{HV}]\, N^{-1}_{HV\text{-max}} \ , \end{cases}$$

where $N_{LV\text{-}min}$ and $N_{HV\text{-}max}$ are the numbers of examinees who belong to the lower subgroup and obtained the response pattern V-min , and those who belong to the upper subgroup and obtained the response pattern V-max , respectively.

It can be seen that $\hat{\theta}^*_{V\text{-}min}$ and $\hat{\theta}^*_{V\text{-}max}$, which were defined in the preceding paragraph, are consistent, or converge in probability to $\theta^*_{V\text{-}min}$ and $\theta^*_{V\text{-}max}$, respectively, as the sample sizes increase. In other words, if N_L , N_H , $N_{LV\text{-}min}$ and $N_{HV\text{-}max}$ are large enough, the probabilities with which $\hat{\theta}^*_{V\text{-}min}$ and $\hat{\theta}^*_{V\text{-}max}$ assume values within the vicinities of $\theta^*_{V\text{-}min}$ and $\theta^*_{V\text{-}max}$, respectively, will be very high. Although the two numbers, $N_{LV\text{-}min}$ and $N_{HV\text{-}max}$, also depend upon the choice of the interval, $(\underline{\theta},\bar{\theta})$, by virtue of the Monte Carlo method, we can control the two sample sizes, N_L and N_H , as we wish.

A procedure wich which we may obtain $\hat{\theta}^*_{V\text{-}min}$ and $\hat{\theta}^*_{V\text{-}max}$, which are defined by (6.11), can be summarized as follows.

(1) Determine the interval, $(\underline{\theta},\bar{\theta})$.

(2) Obtain the critical value, θ_c .

(3) Determine the sample size, N , which makes both N_L and N_H large enough for our purpose.

(4) Produce the ability levels of these N hypothetical subjects from the uniform distribution for the interval, $(\underline{\theta},\bar{\theta})$. This can be done either by the Monte Carlo method, or by placing the N examinees at the equally spaced points in the entire interval, $(\underline{\theta},\bar{\theta})$, or using one of its variations.

(5) Calibrate by the Monte Carlo method a response pattern for each of the N hypothetical examinees with respect to the n test items of our test.

(6) Find out the two frequencies, N_{LV} and N_{HV} , for each

response pattern V .

(7) Obtain the maximum likelihood estimate $\hat{\theta}_V$ for each response pattern whose frequencies, N_{LV} and N_{HV} , are not both zero, excluding V-min and V-max .

(8) Use the above results in (6.11), and compute $\hat{\theta}^*_{V\text{-min}}$ and $\hat{\theta}^*_{V\text{-max}}$.

Note that the probabilities with which we obtain positive frequencies for $N_{LV\text{-max}}$ and $N_{HV\text{-min}}$ are both negligibly small, and this fact can be used as a checking process.

Thus we can define the new estimator, $\hat{\theta}^*_V$, such that

$$(6.12)\qquad \hat{\theta}^*_V \begin{cases} = \hat{\theta}^*_{V\text{-min}} & \text{for } V = V\text{-min} \\ = \hat{\theta}^*_{V\text{-max}} & \text{for } V = V\text{-max} \\ = \hat{\theta}_V & \text{otherwise,} \end{cases}$$

as distinct from θ^*_V , which is defined by (6.4). Unlike θ^*_V , this estimator, $\hat{\theta}^*_V$, depends upon the Monte Carlo method, and, therefore, it has some fluctuations. In order to reach high accuracies, we need large numbers for N_L and N_H .

(VI.4) Bayes Estimator with a Uniform Density as the Prior

Let μ'_{1V} be the Bayes estimator with the prior, $f_V(\theta)$. We can write

$$(6.13)\qquad \mu'_{1V} = \int_{-\infty}^{\infty} \theta\, f_V(\theta)\, d\theta \quad ,$$

where $f_V(\theta)$ is the density function of θ for the subgroup of examinees whose response patterns are uniformly V , which is given by

$$(6.14)\qquad f_V(\theta) = f(\theta)\, P_V(\theta)\, [\int_{-\infty}^{\infty} f(\theta)\, P_V(\theta)\, d\theta]^{-1} \ .$$

This estimator is the one which makes the mean square error, such that

$$(6.15)\qquad Q = E[(\theta_V^{**} - \theta)^2] \ ,$$

minimal (Samejima, 1969), where θ_V^{**} is any conceivable estimator of θ based upon the response pattern V . It is obvious that this estimator heavily depends upon the prior.

We can think of a population-free estimator based on the Bayes estimator, by removing the influence of a particular prior. Let us assume that we can more or less specify the interval, $(\underline{\theta},\bar{\theta})$, for which our test is meaningful. To lift the effect of a given prior, we shall use the uniform density for this interval of θ . Let $\mu_{1V}^{*\prime}$ be the resultant estimator. Thus we have

$$(6.16)\qquad f(\theta) \begin{cases} = (\bar{\theta} - \underline{\theta})^{-1} & \text{for } \underline{\theta} \leqslant \theta \leqslant \bar{\theta} \\ = 0 & \text{otherwise.} \end{cases}$$

Substituting (6.16) into (6.13) and rearranging, we obtain

$$(6.17)\qquad \mu_{1V}^{*\prime} = \int_{\underline{\theta}}^{\bar{\theta}} \theta\, P_V(\theta)\, d\theta\, [\int_{\underline{\theta}}^{\bar{\theta}} P_V(\theta)\, d\theta]^{-1} \ .$$

Note that this estimator depends solely upon the operating characteristic, $P_V(\theta)$, and the interval, $(\underline{\theta},\bar{\theta})$, for which our test is meaningful.

In practice, it may not be wise to use this estimator, since even with a relatively small number of test items the number of response patterns is so large and the calculation of the estimates is time-consuming. We could use two estimates, $\mu_{1V\text{-min}}^{*\prime}$ and

$\mu^{*'}_{1V\text{-max}}$, for the replacement of negative and positive infinities of the maximum likelihood estimate for the two extreme response patterns, V-min and V-max , however, without going through too tedious computations.

(VI.5) Subtest 3

We notice in Figure 3-4-1 that the square root of the test information function for Subtest 3 decreases quickly as θ departs from the middle part of the interval, (-3.0, 3.0) . With this subtest as the Old Test, we find fourteen hypothetical examinees who obtained V-min , and twelve who obtained V-max , for their response patterns. Since this is as large as 5.2 percent of the total number of examinees, instead of excluding them, we decided to keep them and experiment with them on the alternative estimator, $\hat{\theta}^*_V$, which was introduced in the Section VI.3 .

With Subtest 9 as our Old Test, we find one examinee who obtained V-min and one whose response pattern is V-max . In this case, we excluded these two from our original data and used 498 examinees in our estimation process, since there are only two and their exclusion will not change the result substantially. With all the other subtests, none of our hypothetical examinees obtained V-min or V-max .

Table 6-5-1 presents the identification number and the ability level for each of the fourteen hypothetical examinees who obtained V-min and the twelve who obtained V-max , for Subtest 3 . We can see in this table that, although most of these twenty-six examinees have the ability levels equal to or close to one of the two extreme values of θ , -2.475 and 2.475 , there are some examinees, like 118 , 210 and 491 , whose ability levels are substantially less than 2.475 in absolute values. Tables 6-5-2 and 6-5-3 present the response patterns of these twenty-six examinees for the ten unknown, binary test items.

We need some modification to the estimator, however. Since

TABLE 6-5-1

Identification Number and Ability Level of Each of the Fourteen Hypothetical Examinees Who Obtained V-min , and of the Twelve Who Obtained V-max of Subtest 3 .

ID	θ	ID	θ
1	-2.475	491	2.025
101	-2.475	193	2.125
201	-2.475	493	2.125
401	-2.475	294	2.175
2	-2.425	296	2.275
102	-2.425	397	2.325
202	-2.425	98	2.375
302	-2.425	198	2.375
303	-2.375	199	2.425
4	-2.325	299	2.425
108	-2.125	499	2.425
109	-2.075	300	2.475
210	-2.025		
118	-1.625		

the square root of the test information function of Subtest 3 is not constant, we must transform θ to τ in the process of estimating the operating characteristics of the item scores. We recall that, with the transformed scale of ability, the asymptotic unbiasedness and the normality were used as the approximation to the conditional distribution of the maximum likelihood estimate, $\hat{\tau}_V$, given τ . We need, therefore, the unbiasedness of the modified estimator with respect to τ , instead of θ . Let $\hat{\tau}^*_V$ be the estimator with respect to τ . Thus we can write

$$(6.18) \qquad \hat{\tau}^*_V \begin{cases} = \hat{\tau}^*_{V\text{-min}} & \text{for } V = V\text{-min} \\ = \hat{\tau}^*_{V\text{-max}} & \text{for } V = V\text{-max} \\ = \hat{\tau}_V & \text{otherwise} \end{cases}$$

where $\hat{\tau}^*_{V\text{-min}}$ and $\hat{\tau}^*_{V\text{-max}}$ are defined by

TABLE 6-5-2

Identification Number and the Response Pattern of the Ten Unknown, Binary Items Obtained by Each of the Fourteen Hypothetical Examinees Whose Response Patterns of Subtest 3 are V-min .

ID	Response Pattern
1	0001000000
101	0100000000
201	0100000000
401	1000000000
2	0100000000
102	0000000000
202	0000000000
302	1000000000
303	1000000000
4	1100000000
108	1000000000
109	1001000000
210	1000000000
118	1010000000

$$(6.19)\quad \begin{cases} \hat{\tau}^{*}_{V\text{-min}} = [\frac{1}{2}(\tau_c + \underline{\tau})\ N_L - \sum\limits_{\substack{V\neq V\text{-min}\\ V\neq V\text{-max}}} \hat{\tau}_V\ N_{LV}]\ N^{-1}_{LV\text{-min}} \\ \hat{\tau}^{*}_{V\text{-max}} = [\frac{1}{2}(\underline{\tau} + \tau_c)\ N_H - \sum\limits_{\substack{V\neq V\text{-min}\\ V\neq V\text{-max}}} \hat{\tau}_V\ N_{HV}]\ N^{-1}_{HV\text{-max}} \end{cases} ,$$

In these formulas, τ_c , $\underline{\tau}$, $\bar{\tau}$, and the maximum likelihood estimate, $\hat{\tau}_V$, can uniformly be transformed from θ_c , $\underline{\theta}$, $\bar{\theta}$ and $\hat{\theta}_V$, by means of $\tau = \tau(\theta)$.

Figure 6-5-1 presents the two operating characteristics, $P^{*}_{V\text{-min}}(\tau)$ and $P^{*}_{V\text{-max}}(\tau)$, as functions of τ , by solid and dotted curves, respectively. In the same figure, also presented are the positions of two τ_c 's which we used separately. Eight different intervals were used for $(\underline{\tau},\bar{\tau})$, and the results are

TABLE 6-5-3

Identification Number and the Response Pattern of the Ten Unknown, Binary Items Obtained by Each of the Twelve Hypothetical Examinees Whose Response Patterns of Subtest 3 are V-max .

ID	Response Pattern
491	1111111000
193	1111111110
493	1111111110
294	1111111111
296	1111111111
397	1111111111
98	1111111111
198	1111101110
199	1111111111
299	1111111110
499	1111111111
300	1111111111

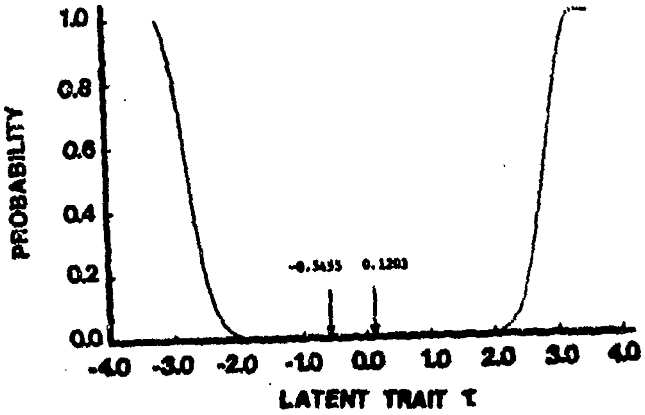


FIGURE 6-5-1

Operating Characteristics of V-min (Solid Line) and V-max (Dotted Line) of Subtest 3 Given As Functions of the Transformed Latent Trait τ , Together with the Critical Value , τ_c , Set at Two Different Positions.

shown in Tables 6-5-4 and 6-5-5 for $\tau_c = 0.1203$ and $\tau_c = -0.5455$, respectively. It is obvious that for the first three intervals of

TABLE 6-5-4

Two Estimates, $\hat{\tau}^*_{V-min}$ and $\hat{\tau}^*_{V-max}$, Obtained by Using Each of the Eight Different Intervals, $(\underline{\tau}, \bar{\tau})$, and $\tau_c = 0.1203$ for Subtest 3. The Sample Sizes, N_L, N_H and N, Together with the Two Frequencies N_{V-min} and N_{V-max}, Are Also Presented for Each Case.

Case	$\underline{\tau}$	$\bar{\tau}$	$\hat{\tau}^*_{V-min}$	$\hat{\tau}^*_{V-max}$	N_{V-min}	N_{V-max}	N_L	N_H	N
1	-1.8456	2.0771	2.9707	-0.6316	1	3	1,640	1,630	3,270
2	-2.0521	2.2668	5.8168	0.6564	1	10	1,810	1,790	3,600
3	-2.2461	2.4373	-1.5891	1.7371	8	19	1,970	1,930	3,900
4	-2.4273	2.5860	-1.8162	2.2439	23	32	2,125	2,055	4,180
5	-2.5131	2.6516	-2.2006	2.4000	39	42	2,195	2,110	4,305
6	-2.6757	2.7636	-2.5467	2.6242	81	74	2,330	2,205	4,535
7	-2.8267	2.8095	-2.7265	2.7370	145	93	2,455	2,240	4,695
8	-3.0000	3.0000	-2.8432	2.8855	258	196	2,600	2,400	5,000

TABLE 6-5-5

Two Estimates, $\hat{\tau}^*_{V-min}$ and $\hat{\tau}^*_{V-max}$, Obtained by Using Each of the Eight Different Intervals, $(\underline{\tau}, \bar{\tau})$, and $\tau_c = -0.5455$ for Subtest 3. The Sample Sizes, N_L, N_H and N, Together with the Two Frequencies N_{V-min} and N_{V-max}, Are Also Presented for Each Case.

Case	$\underline{\tau}$	$\bar{\tau}$	$\hat{\tau}^*_{V-min}$	$\hat{\tau}^*_{V-max}$	N_{V-min}	N_{V-max}	N_L	N_H	N
1	-1.8456	2.0771	7.7998	-2.2507	1	3	1,085	2,185	3,270
2	-2.0521	2.2668	11.3745	0.1132	1	10	1,255	2,345	3,600
3	-2.2461	2.4373	-0.8183	1.4841	8	19	1,415	2,485	3,900
4	-2.4273	2.5860	-1.6061	2.0856	23	32	1,570	2,610	4,180
5	-2.5131	2.6516	-2.0651	2.2750	39	42	1,640	2,665	4,305
6	-2.6757	2.7636	-2.4788	2.5455	81	74	1,775	2,760	4,535
7	-2.8267	2.8095	-2.6867	2.6865	145	93	1,900	2,795	4,695
8	-3.0000	3.0000	-2.8214	2.8596	258	196	2,045	2,955	5,000

τ the results are meaningless, since the two frequencies, $N_{V\text{-min}}$ and $N_{V\text{-max}}$, are so small. We can also see that, as these frequencies grow larger, the resultant estimates get closer to each other over the two different values of τ_c.

To compare these results with the largest and the smallest finite maximum likelihood estimates for Subtest 3, the values of $\hat{\tau}_V$ for the fifteen response patterns in which only one item is answered correctly, and those for the fifteen other response patterns in which only one item is answered incorrectly, are shown in Table 6-5-6. We can see in this table that the least finite $\hat{\tau}_V$ is -2.6518 and the greatest finite $\hat{\tau}_V$ is 2.7683. We notice in Tables 6-5-4 and 6-5-5 that only the largest interval of τ, (-3.0, 3.0), provides us with two alternative estimates, which are greater in absolute values than those two finite estimates. Our selection is, therefore, -2.843 for $\hat{\tau}^*_{V\text{-min}}$, and 2.885 for

TABLE 6-5-6

Fifteen Response Patterns of Subtest 3, Each of Which Consists of Fourteen Zeros and One "1", and the Corresponding Two Maximum Likelihood Estimates, $\hat{\theta}_V$ and $\hat{\tau}_V$, for Each Response Pattern, and Another Set of Fifteen Response Patterns, Each of Which Has (n-1) m_g's and One (m_g-1) and the Corresponding $\hat{\theta}_V$ and $\hat{\tau}_V$ for Each.

Response Pattern	$\hat{\theta}_V$	$\hat{\tau}_V$	Respouse Pattern	$\hat{\theta}_V$	$\hat{\tau}_V$
000000000000001	-1.3998	-1.7296	222222222222221	2.3526	2.6855
000000000000010	-1.5206	-1.8562	222222222222212	2.3454	2.6800
000000000000100	-1.9182	-2.2347	222222222222122	2.4651	2.7683
000000000001000	-1.6990	-2.0336	222222222221222	2.2762	2.6258
000000000010000	-1.9465	-2.2592	222222222212222	2.3359	2.6727
000000000100000	-1.8783	-2.1995	222222222122222	2.1981	2.5620
000000001000000	-1.8346	-2.1603	222222221222222	2.0525	2.4359
000000010000000	-2.0033	-2.3075	222222212222222	2.0810	2.4613
000000100000000	-2.0205	-2.3218	222222122222222	1.9725	2.3627
000001000000000	-2.1792	-2.4483	222221222222222	2.0237	2.4098
000010000000000	-2.0811	-2.3714	222212222222222	1.7479	2.1437
000100000000000	-2.3846	-2.5959	222122222222222	2.0530	2.4363
001000000000000	-2.3887	-2.5987	221222222222222	1.9407	2.3329
010000000000000	-2.3585	-2.5782	212222222222222	1.7595	2.1555
100000000000000	-2.4698	-2.6518	122222222222222	1.8532	2.2488

$\hat{\tau}^*_{V-max}$, respectively. The sample regression of $\hat{\tau}^*_V$ on τ for our five hundred observations turned out to be $0.998\tau + 0.001$ (cf. RR-81-2), which is very close to the unbiasedness.

The other two alternative estimates, $\mu^{*'}_{IV-min}$ and $\mu^{*'}_{IV-max}$, which were introduced in the preceding section, were also computed for each of the eight intervals. These values were calculated with respect to τ , instead of θ , and we obtained $\mu^*_{IV-min} = -1.7434$, -1.9286, -2.0965, -2.2464, -2.3143, -2.4364, -2.5402, -2.7527 and $\mu^*_{IV-max} = 1.9980$, 2.1810, 2.3457, 2.4905, 2.5551, 2.6684, 2.7171, 2.7805 , for Cases 1 through 8, respectively. As the interval of τ grows larger, the resultant estimates get closer to the corresponding values of $\hat{\tau}^*_{V-min}$ and $\hat{\tau}^*_{V-max}$. We did not use them as the substitutes for negative and positive infinities of the maximum likelihood estimate, however, since the conditional unbiasedness of our estimate is an important characteristic in our rationale behind the methods and approaches for estimating the operating characteristics of unknown test items.

(VI.6) Nine Subtests As Our Old Test

In the first year of the present research, the original Old Test was solely used as our Old Test in estimating the item characteristic functions of the ten unknown, binary test items. Thus the first seven research reports, RR-77-1, RR-78-1 through RR-78-6, out of the total eleven, which are written on the estimation of the operating characteristics, are based upon the original Old Test, while the other four research reports, RR-80-2, RR-80-4, RR-81-2 and RR-81-3, are based upon the nine subtests of the original Old Test (cf. Chapter 2). The original Old Test consists of thirty-five test items of three score categories each, whose item parameters are given in Table 3-4-1 of Section III.4 , with each item following the normal ogive model. Furthermore, it has an approximately constant square root of the test information function, 4.65 , for the interval of ability of our interest. This is an ideal situation, and it also provides us with simpler methods and approaches, in which no transformation of ability θ is needed. This situation can be materialized easily in adaptive testing, which we shall observe in Chapter 7.

On the other hand, it will be meaningful to test the robustners of our methods and approaches of estimating the operating characteristics by using a less than ideal Old Test, i.e., one which has fewer test items and a non-constant square root of the test information function. This experiment, if the result turns out to be supportive, will have a benefit of expanding the applicability of our methods and approaches, since in the paper-and-pencil testing situation most tests do not provide us with constant amounts of test information.

The selection of the test items for each of the nine subtests of our original Old Test is shown in Table 3-4-1, and the square root of the test information function is given in Figure 3-4-1, of Section III.4 . We notice that Subtest 3 is also a subtest of Subtest 1, and Subtest 4 is a subtest of Subtest 2, and all the other five subtests are those of the original Old Test only.

In this experimentation, Simple Sum Procedure of the Conditional P.D.F. Approach (cf. Section V.13) with the Normal Approach Method (cf. Section V.9) was selected as our combination of a method and an approach. The main reason for this selection of the Simple Sum Procedure is its simplicity, which does not require the approximation to the density function of $\hat{\tau}$ with respect to each item score category of each unknown test item, as Bivariate P.D.F. Approach does, nor the weight and the proportion which Weighted Sum Procedure and Proportioned Sum Procedure need, respectively. The main reasons why we selected Normal Approach Method are, again, its simplicity, which requires only the first two conditional moments of τ , given $\hat{\tau}$, and the fact that the criterion item characteristic function had been obtained in the Simple Sum Procedure for each of the unknown test items, and the results obtained by the Normal Approach Method, as well as those obtained by the Pearson System Method and the Two-Parameter Beta Method, respectively, turned out to be practically identical with the criterion item characteristic function (cf. Section V.13). With each of Subtest 1, 2 and 3 as our Old Test, both Degree 3 and 4 Cases were applied. For the other six subtests, however, only Degree 4 Case was adopted. The reason for the exclusion of Degree 3 Case in this later research is that, in all the previous studies, the resultant estimated item characteristic functions obtained in Degree 3 Case turned out to be practically identical with those obtained in Degree 4 Case.

As we have seen in the preceding section, with Subtest 3 as our Old Test, we used the set of modified maximum likelihood estimates, $\hat{\tau}_s^*$ $(s=1,2,\ldots,N)$, as our basic data. With each of the other eight subtests as our Old Test, the set of maximum likelihood estimates, $\hat{\tau}_s$, was used.

This part of the research is partly credited to the conscientious effort by one of the author's assistants, Paul Changas.

(VI.7) Sample Linear Regression of $\hat{\tau}_s$ on τ_s

Figure 6-7-1 presents the scatter diagram of ability θ_s

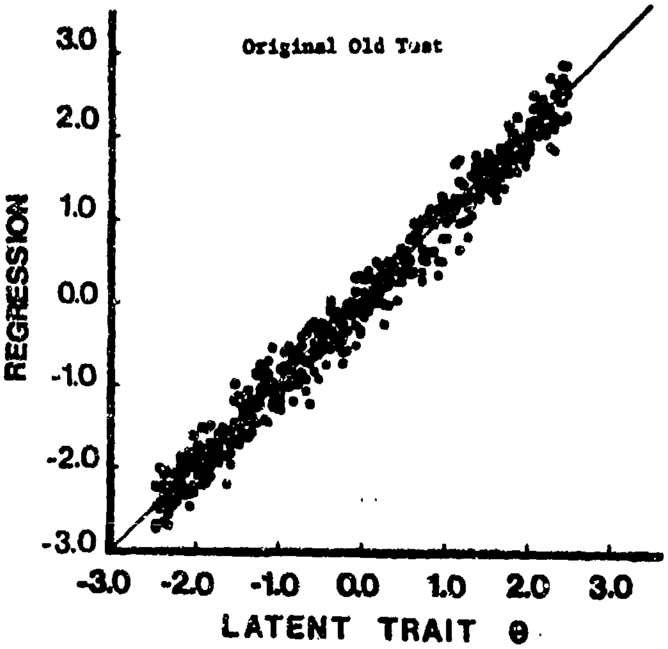


FIGURE 6-7-1

Scatter Diagram of $\hat{\theta}_s$ Plotted Against θ_s for Our Five Hundred Hypothetical Examinees, Which Is Based upon the Original Old Test.

$(s=1,2,\ldots,N)$ and its maximum likelihood estimate, $\hat{\theta}_s$, for our five hundred hypothetical examinees, which are obtained upon our Old Test. We can see in this figure that the conditional unbiasedness of $\hat{\theta}$, given θ , may approximately be satisfied. The sample linear regression of $\hat{\theta}$ on θ , or the best fitted linear function of θ in the least squares sense, turned out to be $1.004\theta - 0.006$, which is very close to the unbiased line, or the linear function which passes the origin with the slope of unity.

Figure 6-7-2 presents the nine scatter diagrams of the transformed latent trait, τ_s $(s=1,2,\ldots,N)$, and its maximum likelihood estimate, $\hat{\tau}_s$, for our five hundred hypothetical examinees, with the exception of the one for Subtest 9, in which four hundred and ninety-eight examinees are used (cf. Section VI.5) . In this figure, for Subtest 3, the modified maximum likelihood estimate, $\hat{\tau}^*_s$, is used instead of the maximum likelihood estimate, $\hat{\tau}_s$. For convenience, we shall not repeat this in the rest of this section and in Section VI.8 , but the reader must understand this is the case. The sample linear regressions of $\hat{\tau}_s$ on τ_s for the seven of the total nine scatter diagrams are as follows.

Subtest 3:	$1.012\tau - 0.004$
Subtest 4:	$1.003\tau + 0.004$
Subtest 5:	$1.018\tau - 0.007$
Subtest 6:	$1.011\tau - 0.000$
Subtest 7:	$1.016\tau - 0.003$
Subtest 8:	$1.000\tau - 0.009$
Subtest 9:	$1.009\tau + 0.013$

We can see that, in all these cases, the sample linear regressions are very close to the unbiasedness line, and practically indistinguishable from it.

Examination of Figure 6-7-2 reveals, however, that the conditional normality of the distribution of $\hat{\tau}$, given τ , may not be approximately satisfied for some subtests. It is obvious that, as the number of test items in the Old Test decreases, the conditional distribution of $\hat{\tau}$, given τ, grows more and more discrete, with the result for Subtest 9 as the climax. Also for some subtests there are some conspicuous diversions from the normality for some range of τ, as we can see in the scatter diagrams for Subtests 2 and 4 in the vicinity of $\tau = 0.0$. We are to see if these deviations from normality visibly affect the resultant estimated

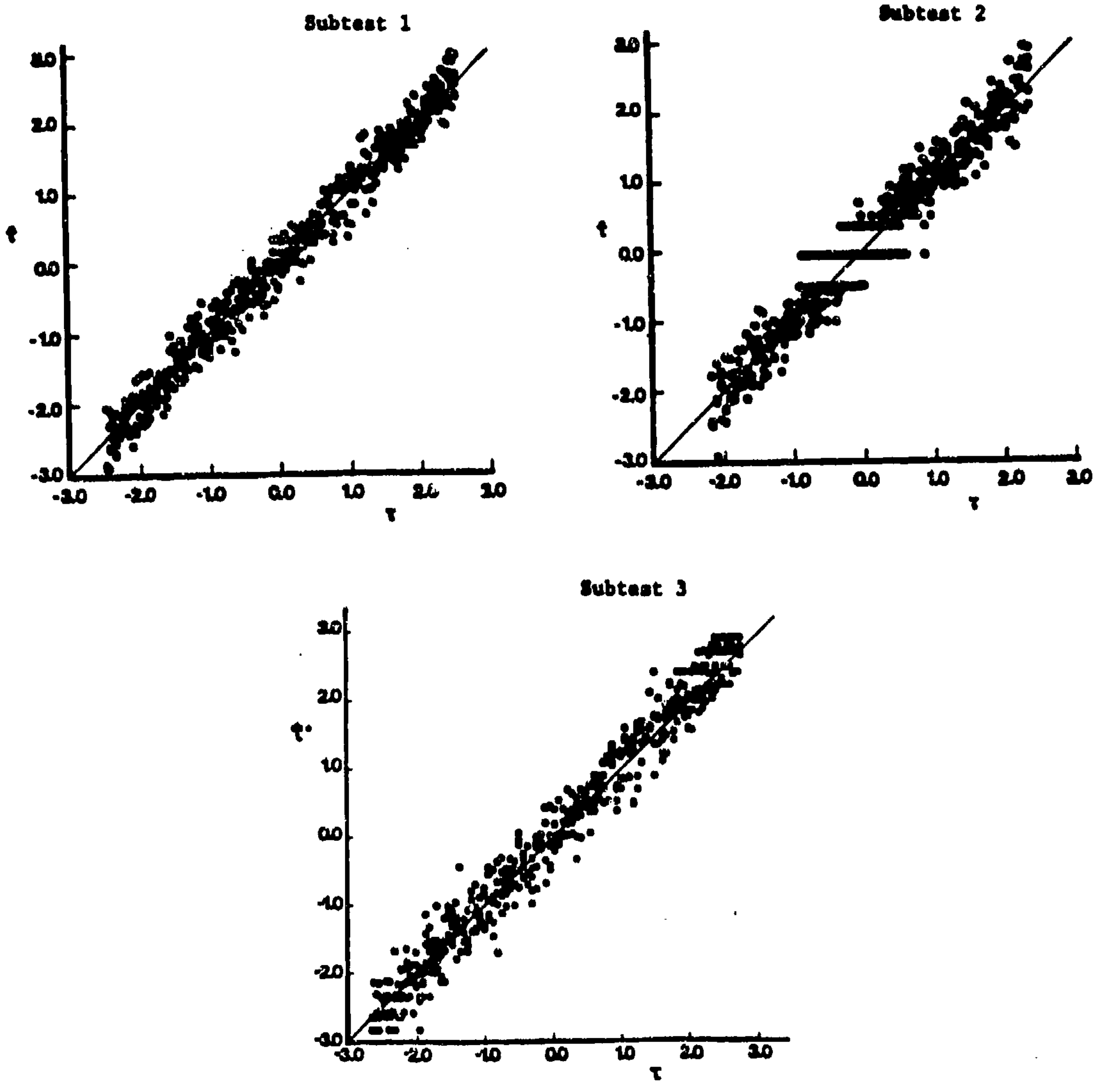
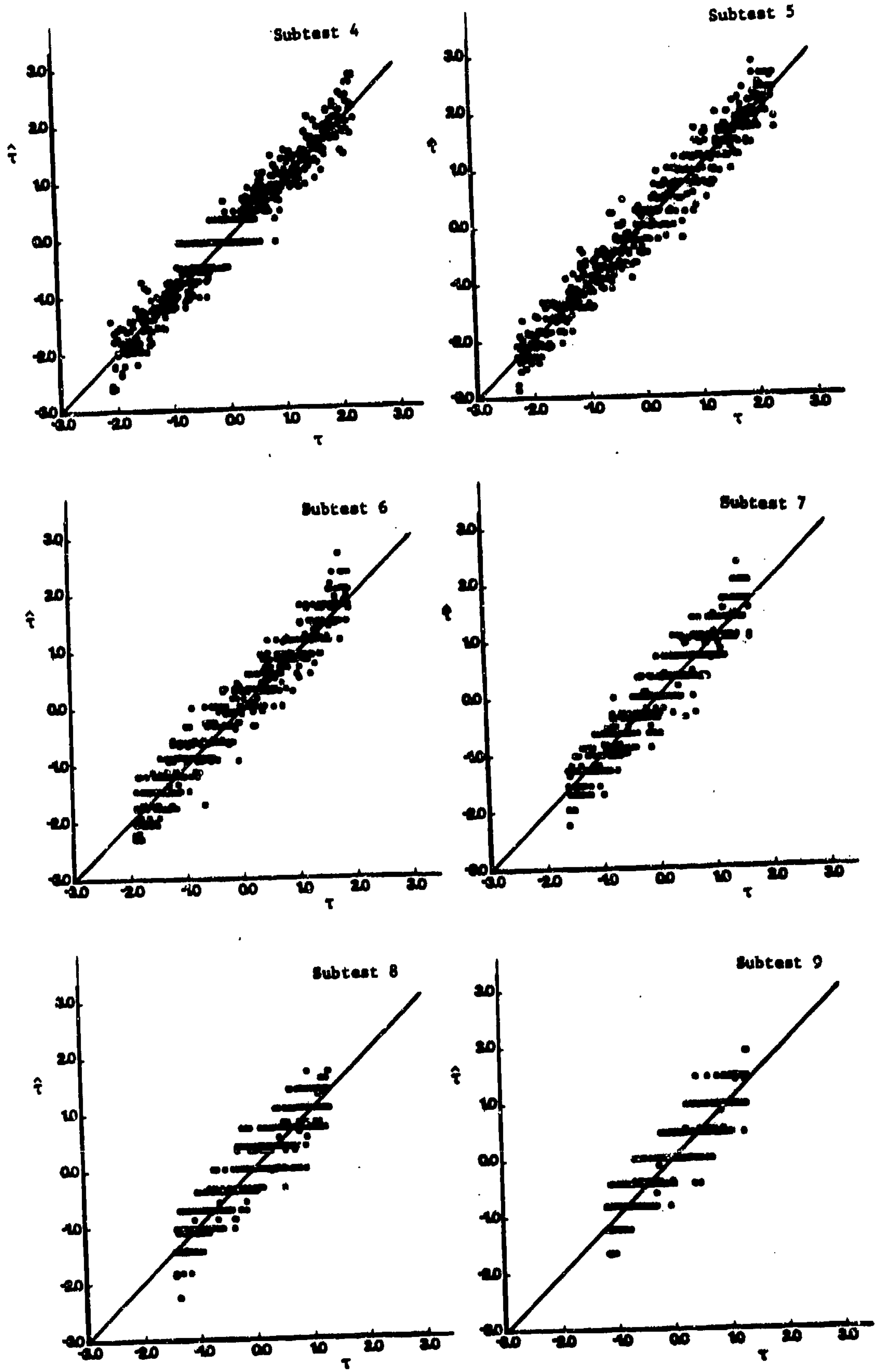


FIGURE 6-7-2

Scatter Diagram of $\hat{\tau}_s$ Plotted Against τ_s for Our Hypothetical Examinees, Which Is Based upon Each Subtest. For Subtest 3, the Estimate Is $\hat{\tau}^*_s$, Instead of $\hat{\tau}$.



Subtest 4

Subtest 5

Subtest 6

Subtest 7

Subtest 8

Subtest 9

FIGURE 6-7-2 (Continued)

item characteristic functions of the unknown binary test items.

(VI.8) Polynomial Approximation to the Density Function, $g(\hat{\tau})$

Figure 6-8-1 presents the two polynomials of degrees 3 and 4 , which were obtained by the method of moments (cf. Section V.6) to approximate the density function, $g(\hat{\tau})$, together with the frequency distribution of the five hundred $\hat{\tau}_s$'s , for each of Subtests 1, 2 and 3 . In each of these three graphs, the resultant polynomial of degree 3 is plotted by a dotted curve, and that of degree 4 is drawn by a solid curve. Approximation to the density function, $g(\hat{\tau})$, by a polynomial was conducted only for Degree 4 Case for each of the other six subtests, the result of which is shown as Figure 6-8-2. We can see in these two figures that there are varieties of different curves and histograms. They are similar for Subtests 2 and 4, but they are not too close for Subtests 1 and 3 , for the latter of which the modified maximum likelihood estimate, $\hat{\tau}^*_s$, was used instead of $\hat{\tau}_s$. The histogram shows greater degrees of ups and downs as the number of test items decreases, the result which was predictable from our observations of the scatter diagrams in the preceding section.

For comparison the reader is suggested to go back to Figure 4-1-2 of Section IV.1 , in which similar graphs are shown for the approximation by the polynomials of degrees 3, 4 and 5, for the five hundred $\hat{\theta}_s$ which were obtained upon the original Old Test.

(VI.9) Estimated Item Characteristic Functions Obtained upon Subtests 1, 2 and 3

As before, for the purpose of illustration, we shall take item 6 as an example. Figure 6-9-1 presents the criterion item characteristics functions (cf. Section V-13) obtained upon Subtests 1, 2 and 3, which are plotted by dotted and short, dashed curves, and dashes and dots, respectively, in comparison with the one obtained upon the original Old Test and the theoretical item characteristic function, which are shown by long, dashed and solid

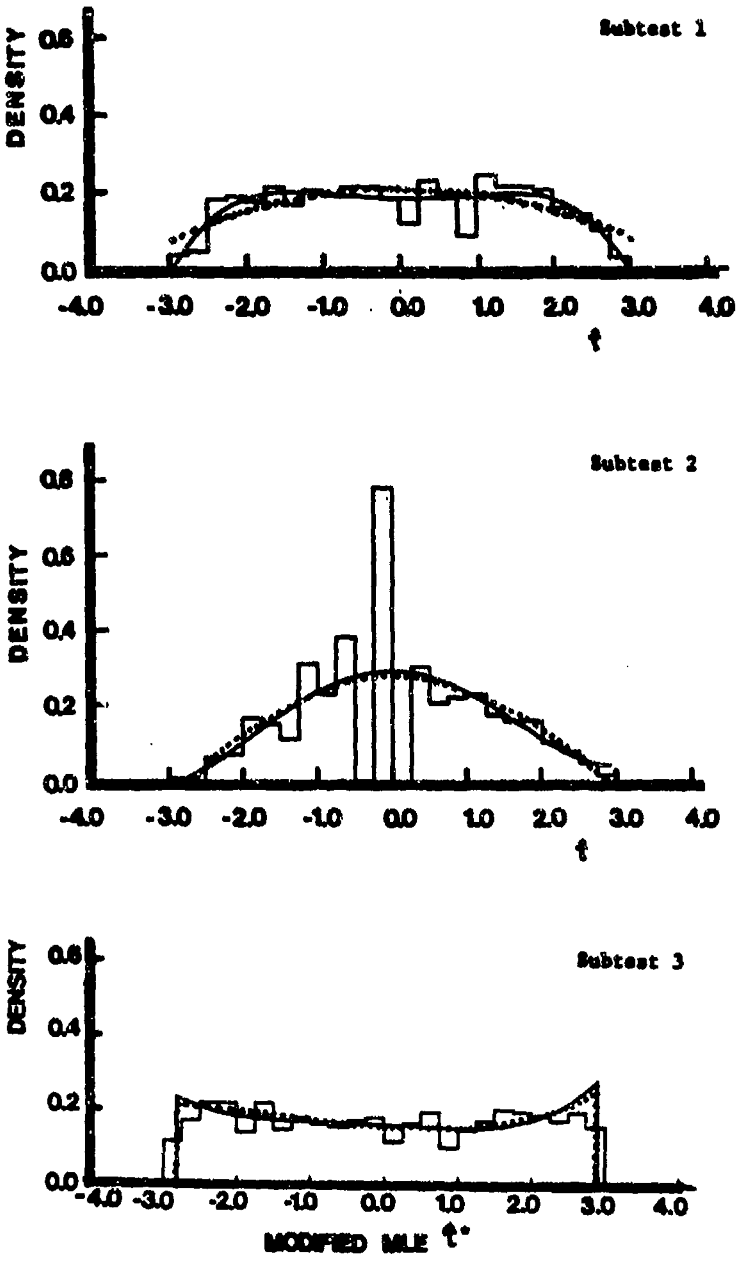


FIGURE 6-8-1

Estimated Density Function, $\hat{g}(\hat{t})$, Obtained by the Method of Moments as a Polynomial of Degree 3 (Dotted Curve) and 4 (Solid Curve), Together with the Relative Frequency Distribution of the Five Hundred $\hat{t}'_s$, for Each of Subtests 1, 2 and 3. For Subtest 3, $\hat{t}^*_s$ is Used Instead of $\hat{t}_s$.

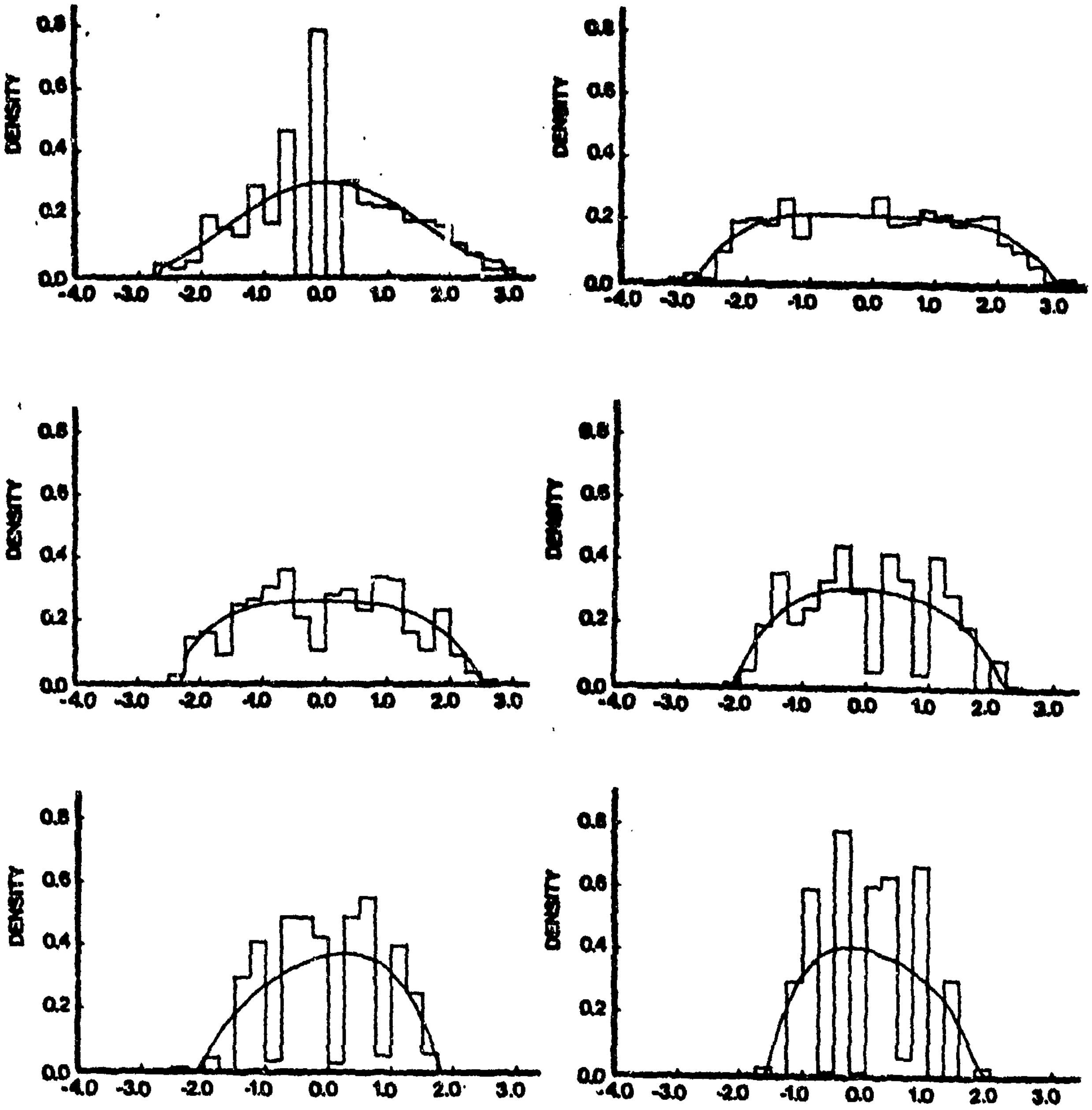


FIGURE 6-8-2

Estimated Density Function, $\hat{g}(\hat{\tau})$, Obtained by the Method of Moments as a Polynomial of Degree 4 , Together with the Relative Frequency Distribution of the Five Hunderd $\hat{\tau}'_s$, for Each of the Six Subtests.

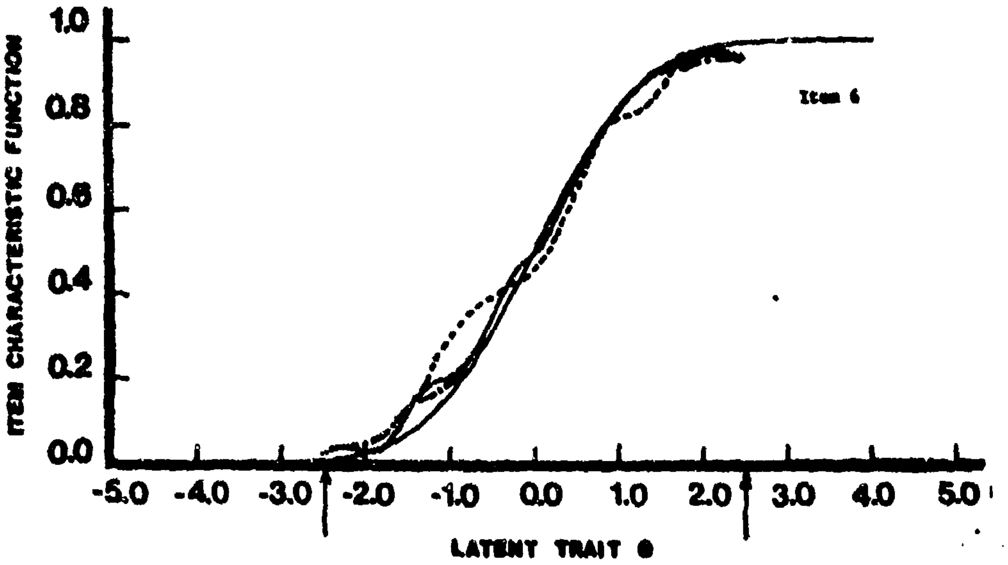


FIGURE 6-9-1

Four Criterion Item Characteristic Functions Obtained upon the Original Old Test (Long, Dashed Curve), upon Subtest 1 (Dotted Curve), upon Subtest 2 (Short, Dashed Curve) and upon Subtest 3 (Dashes and Dots), Together with the Theoretical Item Characteristic Function.

curves. We can see that the two criterion item characteristic functions, which were obtained upon Subtest 3 and upon the original Old Test, are practically indistinguishable, and the one for Subtest 3 is also very close to them. This is a common tendency among all the ten binary test items. In contrast to them, for the interval of θ , (-1.3, 1.6) , the one obtained upon Subtest 2 is substantially different from the other three, and the fitness to the theoretical item characteristic function is a little poorer. This is not the case with all the other nine binary test items, however. In fact, although for items 3, 5, 6 and 7 the fitness is poorer for the ones obtained upon Subtest 2, the order is reversed for items 1, 2, 4, 8 and 10 . It is interesting to note that for items with intermediate difficulty like items 5, 6 and 7 the criterion item characteristic functions fit rather poorly to the corresponding theoretical item characteristic functions. This result is more or less expected from the small amount of test information of Subtest 2 in the vicinity of θ = 0.0 .

Figures 6-9-2 and 6-9-3 present the resultant estimated item characteristic functions of item 6 which are based upon Subtests 1 and 2, respectively, by dotted curves, in both Degree 3 and 4 Cases,

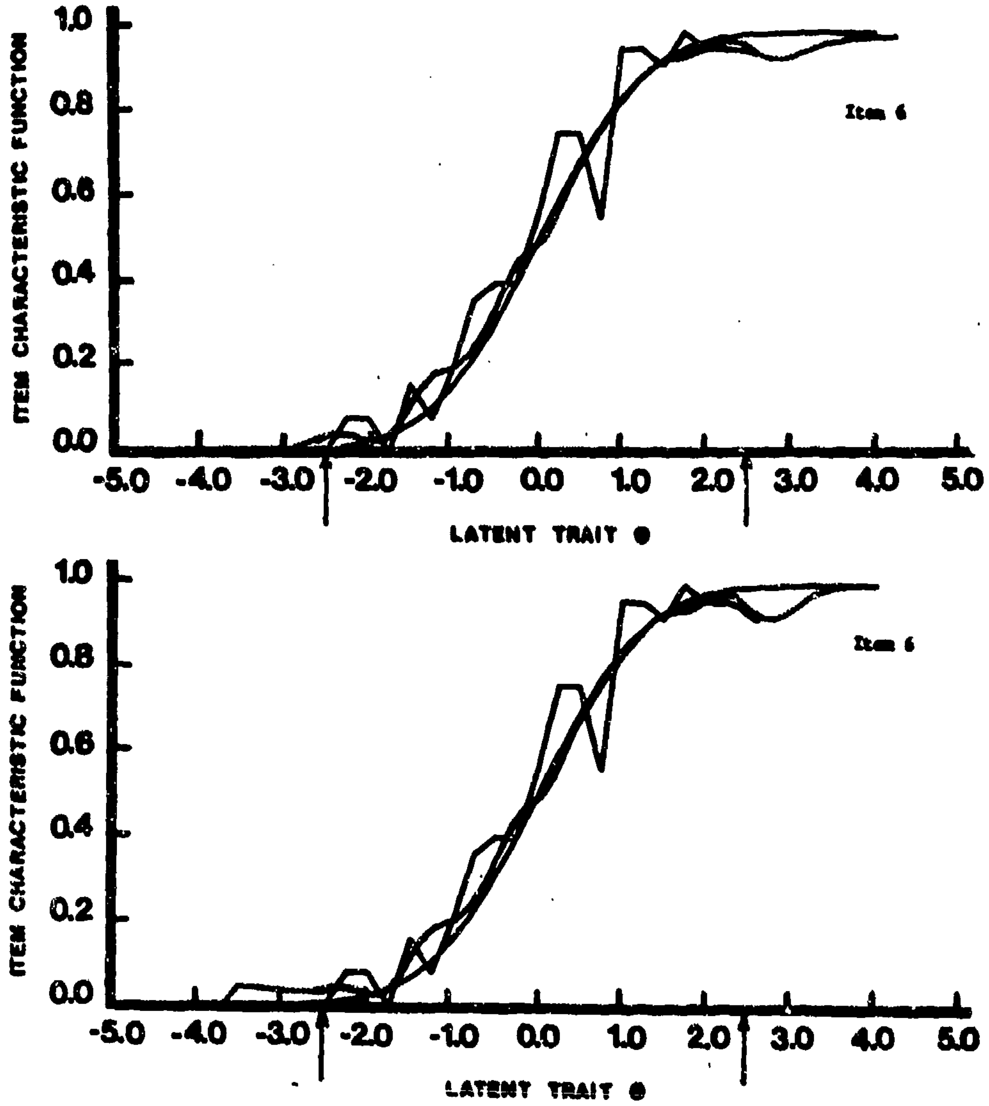


FIGURE 6-9-2

Estimated Item Characteristic Function of Item 6 Based upon Subtest 1 (Dotted Curve) and the One Based upon the Original Old Test (Dashed Curve) Obtained by the Simple Sum Procedure of the Conditional P.D.F. Approach with the Normal Approach Method , in Degree 3 and 4 Cases, in Comparison with the Theoretical Item Characteristic Function (Solid Curve) and the Frequency Ratios of Those Who Answered Correctly (Jagged Solid Line).

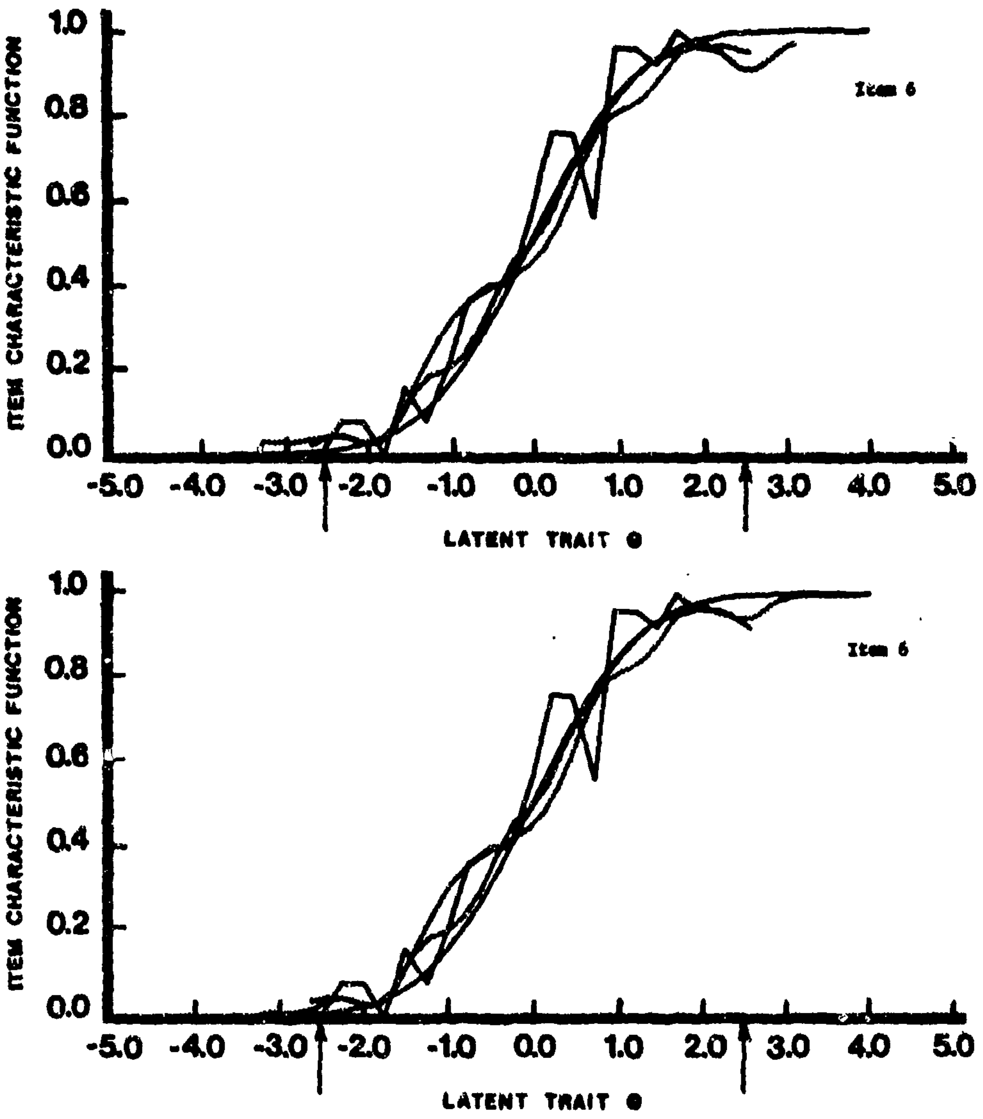


FIGURE 6-9-3

Estimated Item Characteristic Function of Item 6 Based upon Subtest 2 (Dotted Curve) and the One Based upon the Original Old Test (Dashed Curve) Obtained by the Simple Sum Procedure of the Conditional P.D.F. Approach with the Normal Approach Method, in Degree 3 and 4 Cases, in Comparison with the Theoretical Item Characteristic Function (Solid Curve) and the Frequency Ratios of Those Who Answered Correctly (Jagged Solid Line).

together with the corresponding results obtained upon the original Old Test, which are plotted by dashed curves. In the same figures, also presented are the theoretical item characteristic function of item 6, and the frequency ratios of those who answered correctly, by solid curves and jagged solid lines, respectively. It is striking to note that these results in Degree 3 and 4 Cases are practically identical, for the interval of θ , (-2.2, 2.2) , for both Subtests 1 and 2 . We also notice that they are very close to the corresponding criterion item characteristic functions, which we have observed in Figure 6-9-1. These findings are not new, but have been observed repeatedly before, in the results obtained upon the original Old Test. The results for Subtest 1 are practically identical with those obtained upon the original Old Test, for the interval of θ , (-2.2, 2.2). These facts are true not only for item 6, but also for each and every one of the ten binary test items.

Figure 6-9-4 presents the corresponding results for Subtest 2 when the square root of the test information function is approximated by three different polynomials using three subintervals, which is shown in Figure 4-6-3 of Section IV.6 . We can see that the resultant estimated item characteristic functions are very similar to those presented in Figure 4-6-2, in both Degree 3 and 4 Cases. This turned out to be true with all the other nine binary test items: the result which indicates that the crude approximation to the square root of the test information by the single polynomial of degree 7, which is shown in Figure 4-6-2, serves just as well as the more precise one obtained by the three different polynomials.

Figure 6-9-5 presents the corresponding results for Subtest 3. We can see in this figure that the resultant estimated item characteristic function obtained upon Subtest 3 is very close to the one obtained upon the original Old Test, in both Degree 3 and 4 Cases. This is a common tendency among all the ten binary test items. The use of the modified maximum likelihood estimate, $\hat{\tau}^*_s$, certainly did not affect negatively the resultant estimated item characteristic functions.

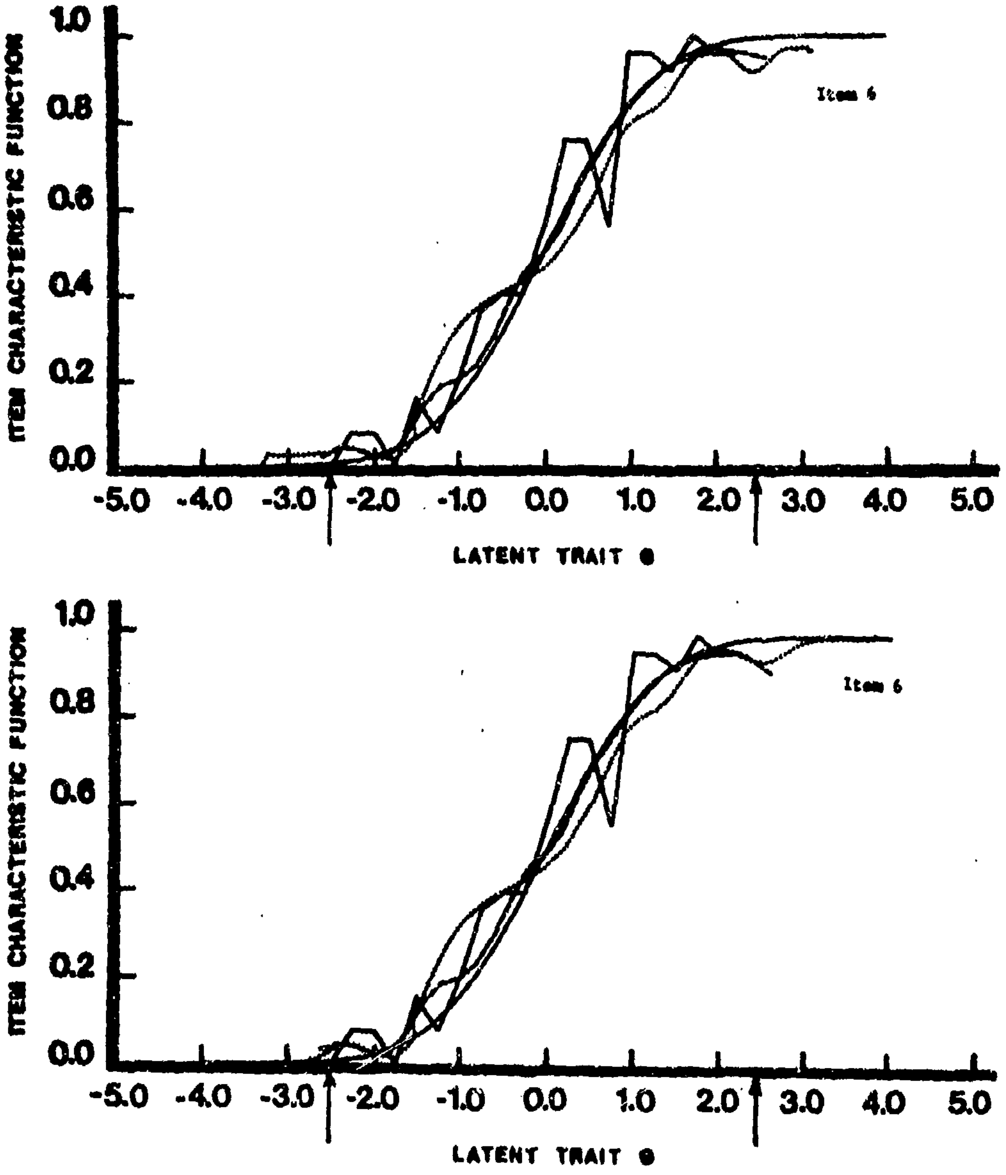


FIGURE 6-9-4.

Estimated Item Characteristic Function of Item 6 Based upon Subtest 2 (Dotted Curve) and the One Based upon the Original Old Test (Dashed Curve) Obtained by the Simple Sum Procedure of the Conditional P.D.F. Approach with the Normal Approach Method, in Degree 3 and 4 Cases, in Comparison with the Theorotical Item Characteristic Function (Solid Curve) and the Frequency Ratios of Those Who Answered Correctly (Jagged Solid Line). The Set of Three Different Polynomials for the Three Subintervals Was Used in the Transformation of θ to τ.

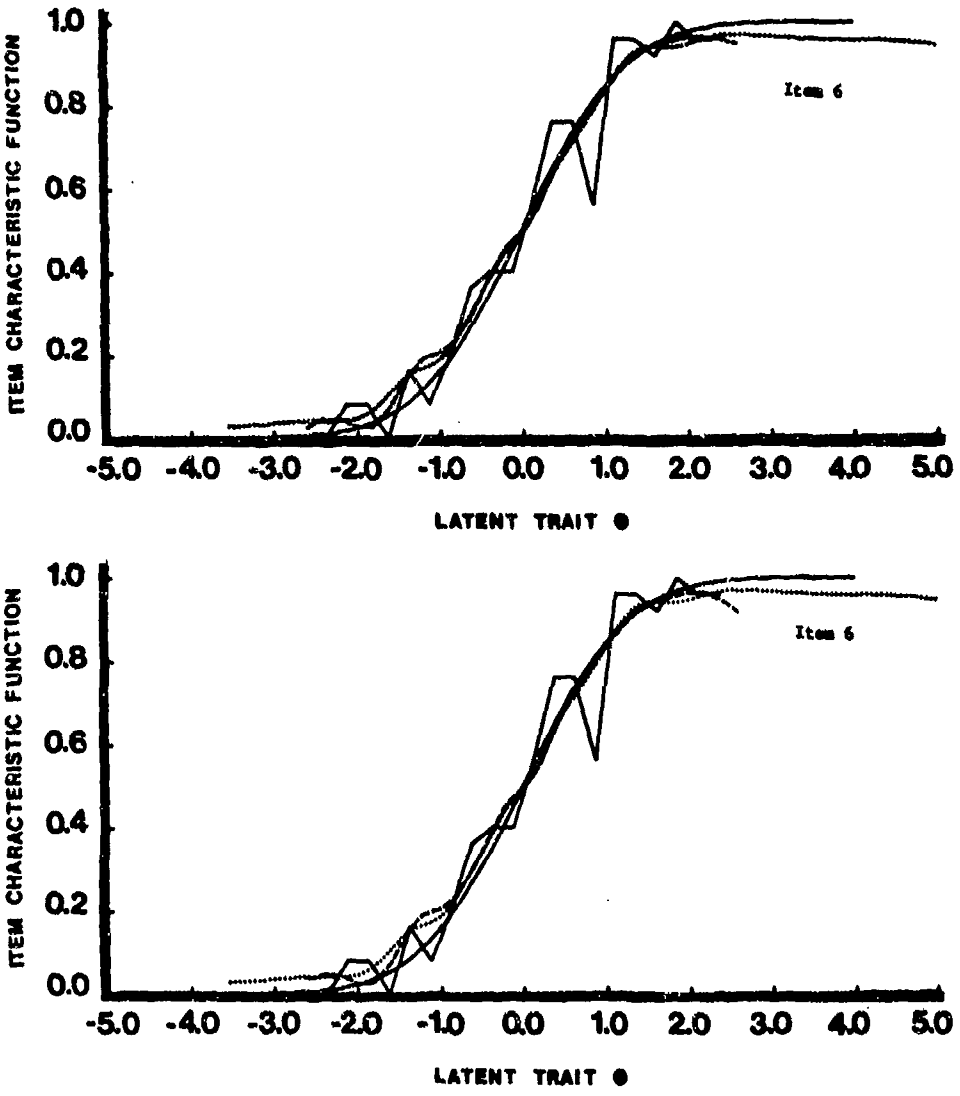


FIGURE 6-9-5

Estimated Item Characteristic Function of Item 6 Based upon Subtest 3 (Dotted Curve) and the One Based upon the Original Old Test (Dashed Curve) Obtained by the Simple Sum Procedure of the Conditional P.D.F. Approach with the Normal Approach Method, in Degree 3 and 4 Cases, in Comparison with the Theoretical Item Characteristic Function (Solid Curve) and the Frequency Ratios of Those Who Answered Correctly (Jagged Solid Line).

(VI.10) Estimated Item Characteristic Functions Obtained upon the Six Other Subtests

Figure 6-10-1 presents the resultant estimated item characteristic functions of item 6 in Degree 4 Case, which were obtained upon Subtests 4, 5, 6, 7, 8 and 9, respectively, by dotted curves, in comparison with the one obtained upon the original Old Test, the theoretical item characteristic function, and the frequency ratios of the correct answer, which are plotted by dashed and solid curves, and jagged solid lines, respectively. We can see in this figure that, up to Subtest 6, the fitness of the resultant estimated item characteristic function to the theoretical item characteristic function is reasonably good, but, after that, it grows flatter. This is a common tendency among all the ten binary test items.

Figure 6-10-2 presents the corresponding results for the other nine binary test items, which were obtained upon Subtest 6. We can see in this figure that the fitness of the estimated item characteristic function is really good for each of these items. In fact, for items 1, 2 and 4 the results fit the corresponding theoretical item characteristic function better than those obtained upon the original Old Test, and they are just as good for items 6, 8 and 10. Considering that Subtest 6 only contains eleven test items, compared with thirty-five in the original Old Test, this result is outstanding. We must conclude, therefore, our combination of a method and an approach is robust over the decrease in number of test items in our Old Test.

It is desirable to experiment on the other combinations of a method and an approach for estimating the operating characteristics, than Simple Sum Procedure of the Conditional P.D.F. Approach with the Normal Approach Method, which we used in the present study. This must wait for future research, however.

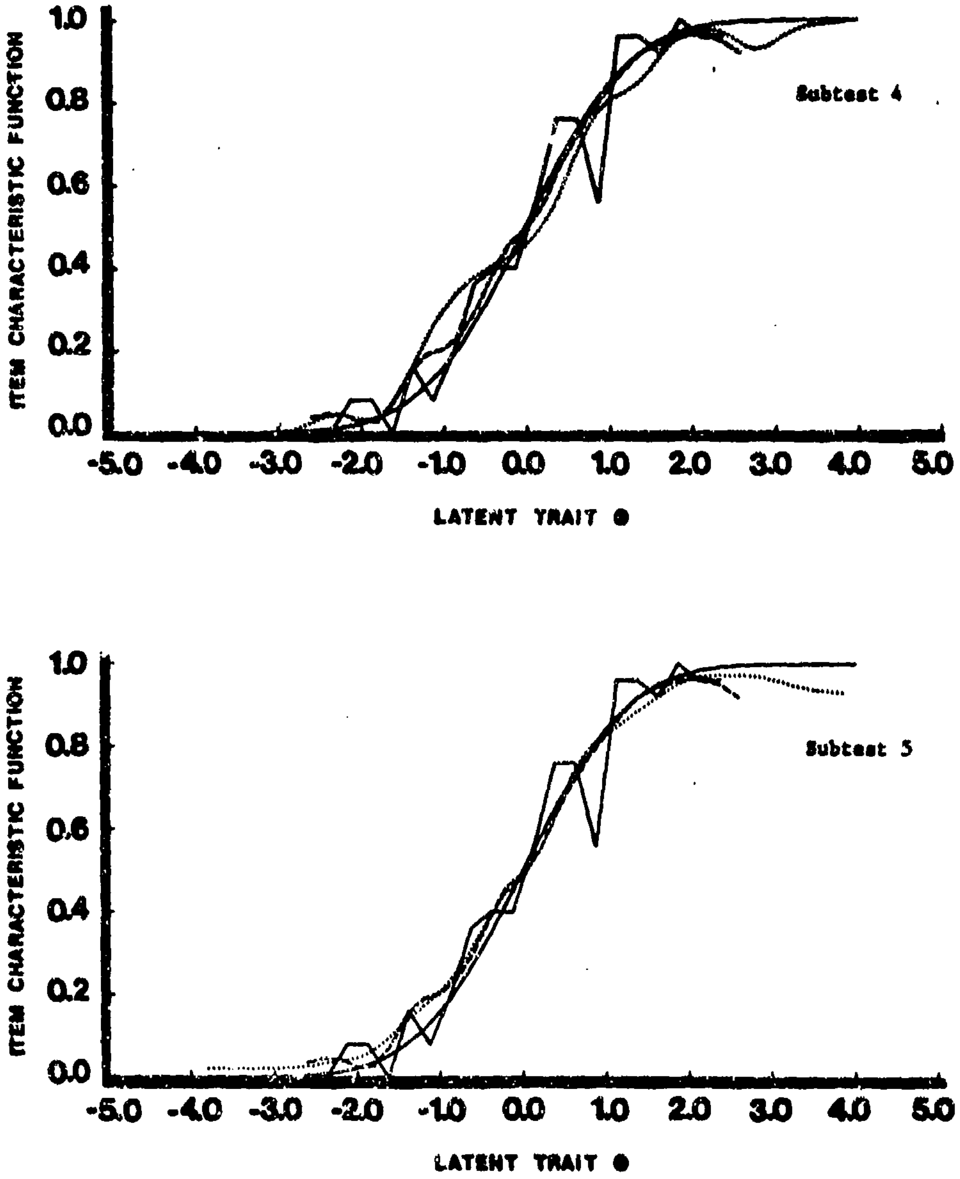
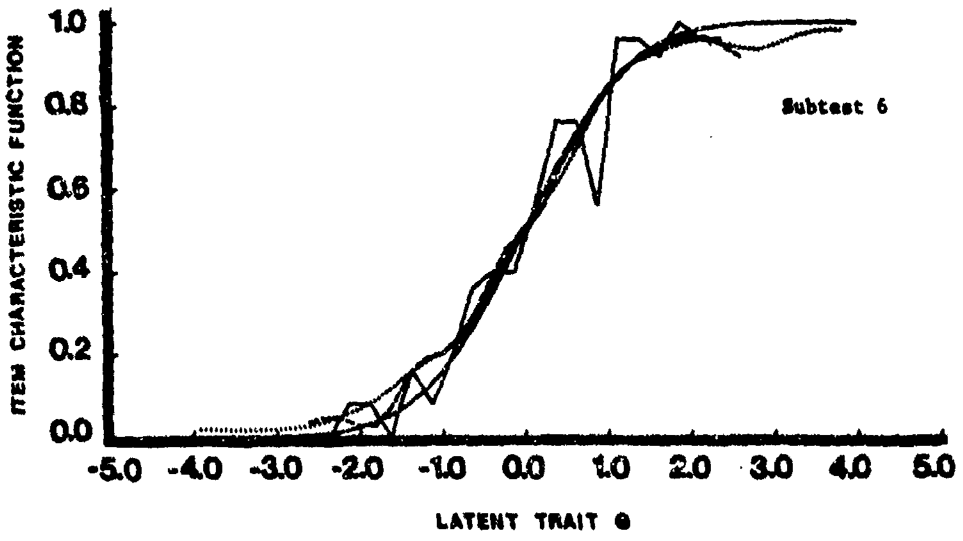


FIGURE 6-10-1

Estimated Item Characteristic Function of Item 6 Based upon Each of Subtests 4 through 9 (Dotted Curve), Obtained by the Simple Sum Procedure of the Conditional P.D.F. Approach and the Normal Approach Method, for Degree 4 Case, in Comparison with the One Based upon the Original Old Test (Dashed Curve), the Theoretical Item Characteristic Function (Smooth Solid Curve) and the Frequency Ratios of Those Who Answered Correctly (Jagged Solid Line).



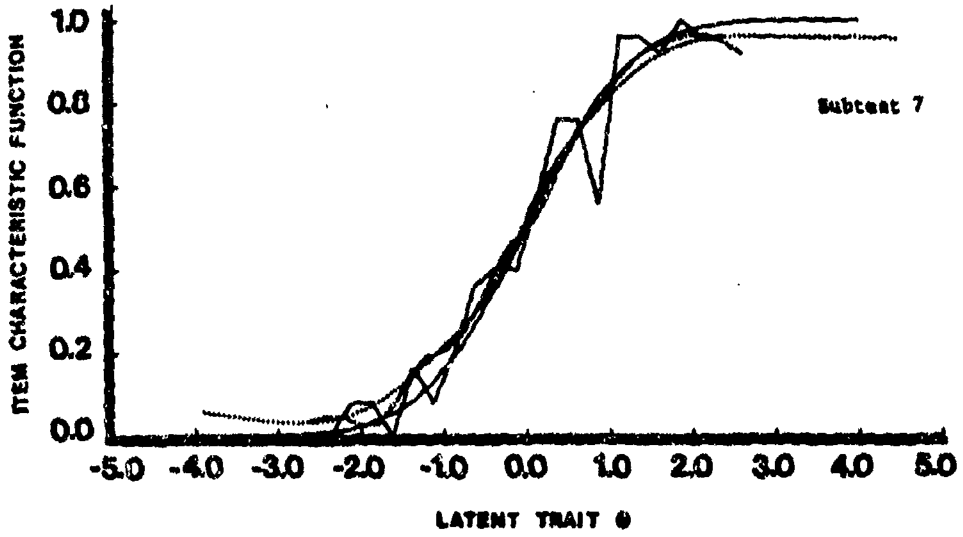
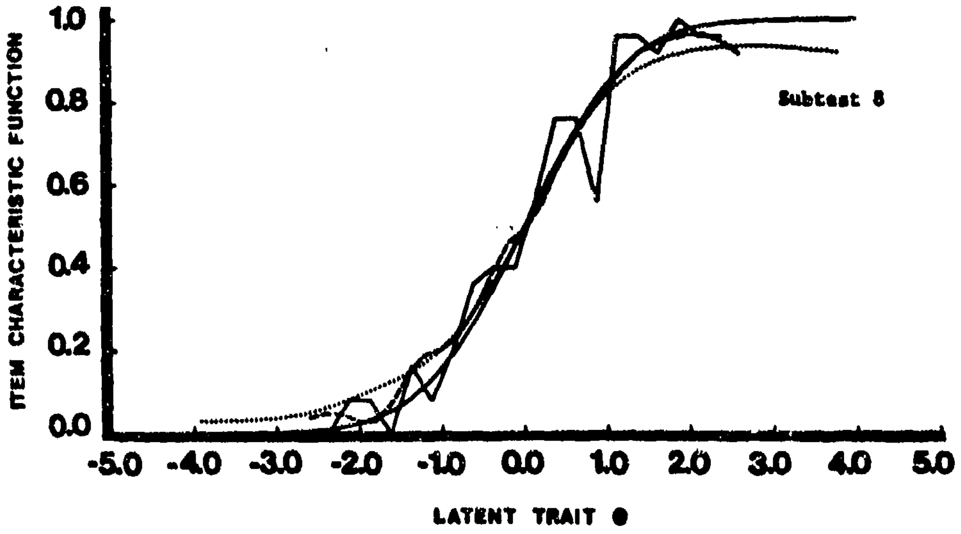


FIGURE 6-10-1 (Continued)



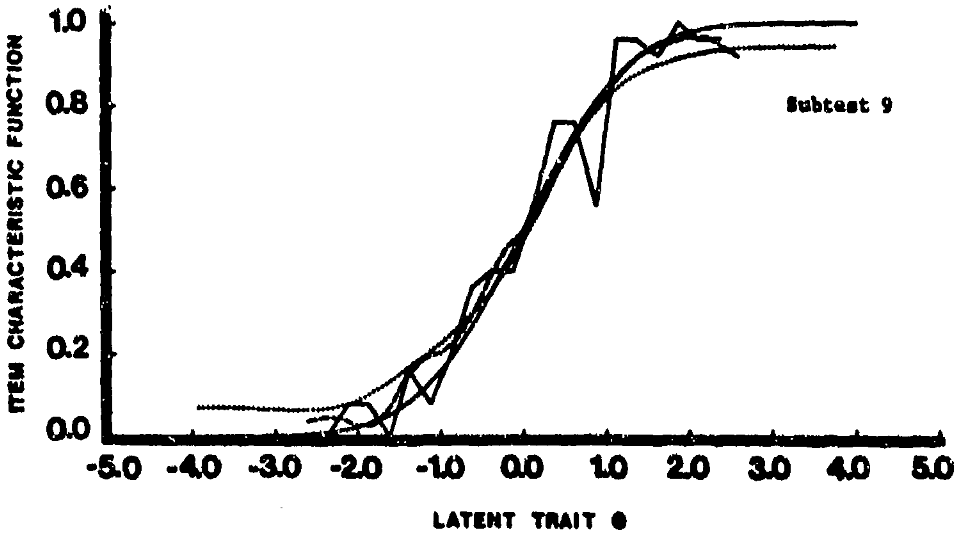


FIGURE 6-10-1 (Continued)

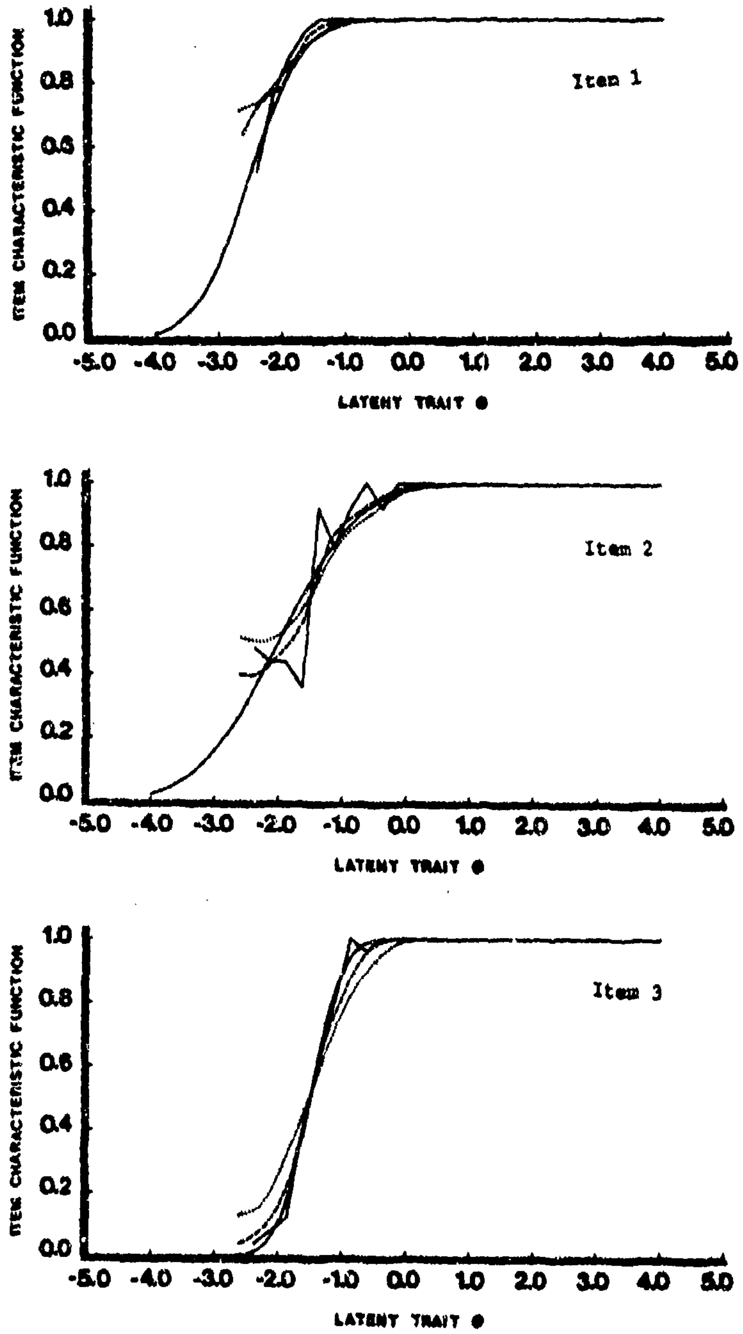
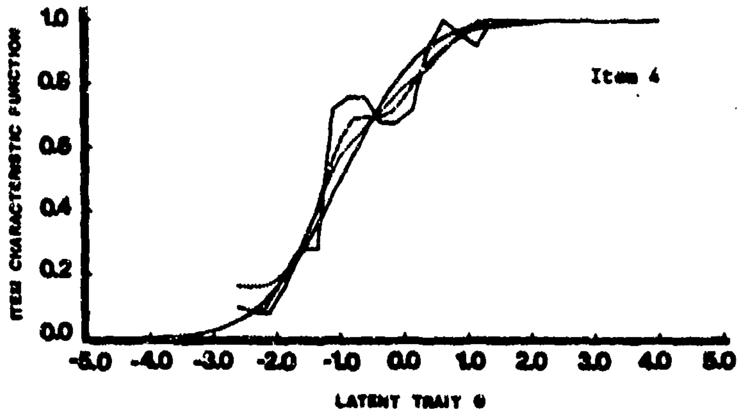
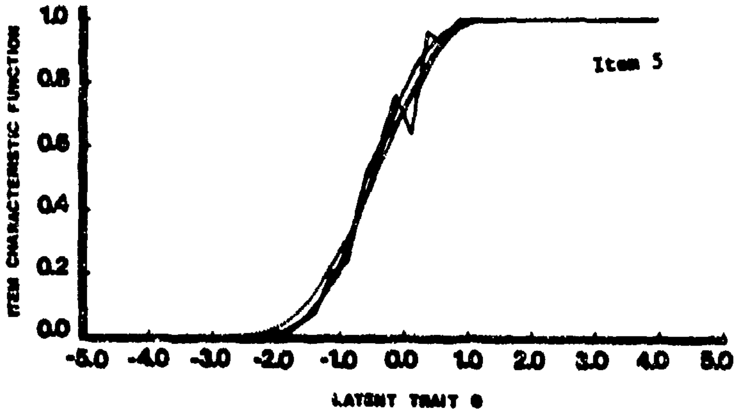


FIGURE 6-10-2

Estimated Item Characteristic Function Based upon Subtest 6 (Dotted Curve) for Each of the Nine Binary Test Items , Which Was Obtained by the Simple Sum Procedure of the Conditional P.D.F. Approach and the Normal Approach Method, for Degree 4 Case , in Comparison with the One Based upon the Original Old Test (Dashed Curve) , the Theoretical Item Characteristic Functiou (Smooth Solid Curve) and the Frequency Ratios of Those Who Answered Correctly (Jagged Solid Line).





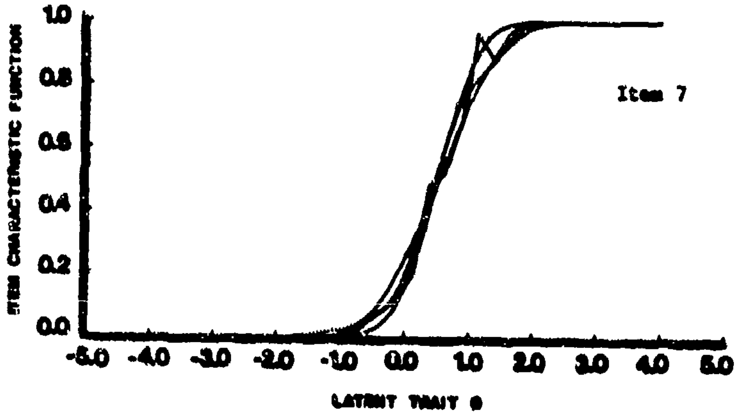


FIGURE 6-10-2 (Continued)

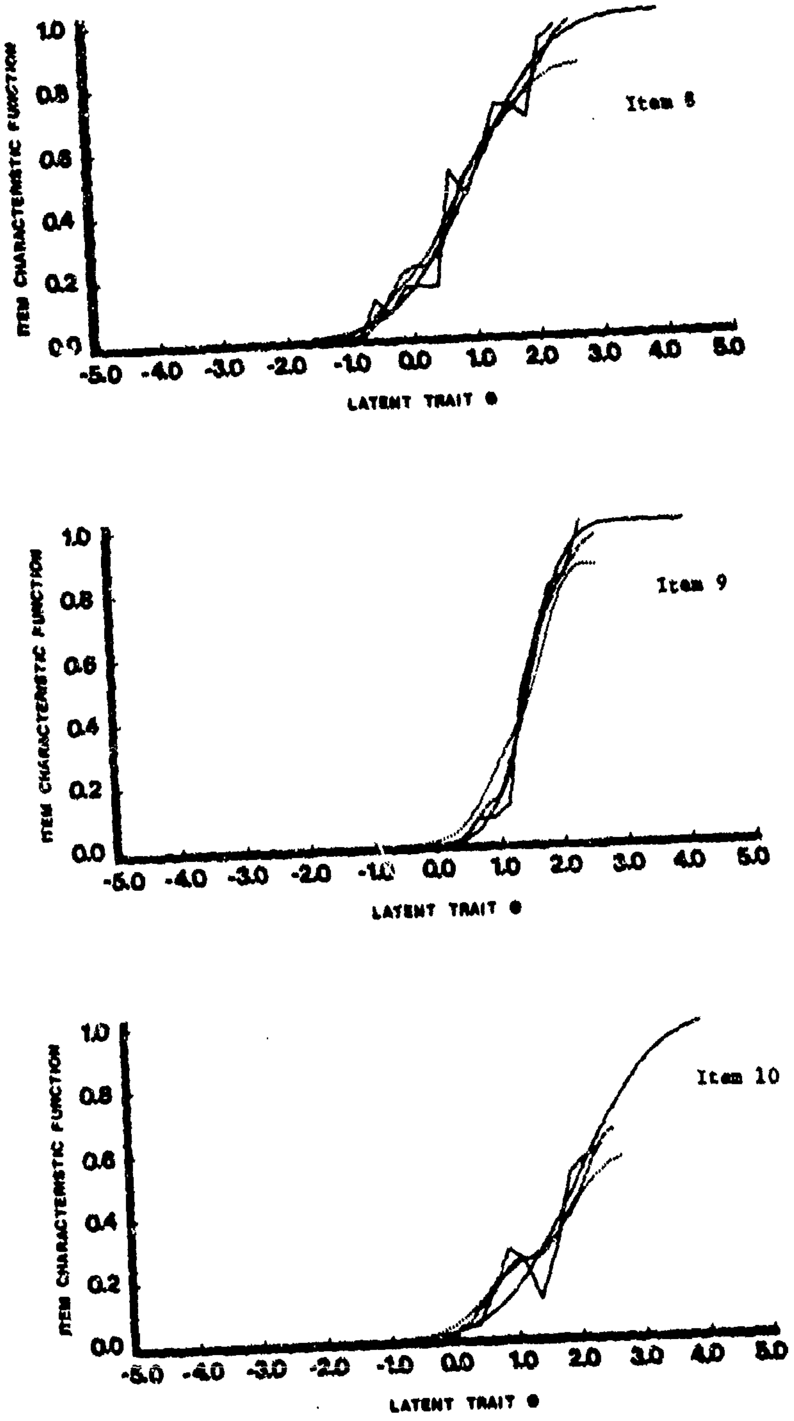


FIGURE 6-10-2 (Continued)

REFERENCES

[1] Indow, T. & Samejima, F. *LIS measurement scale for non-verbal reasoning ability.* Tokyo: Nippon Bunka Kagakusha, 1962. (in Japanese)

[2] Indow, T. & Samejima, F. *On the results obtained by absolute scaling model and the Lord model in the field of intelligence.* Yokohama: Psychological Laboratory, Hiyoshi Campus, Keio University, 1966.

[3] Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17, 1969.

VII Adaptive Testing

In this chapter, we shall observe adaptive testing, or tailored testing, in the context of latent trait theory. By adaptive testing, we mean the testing situation in which test items are selected for an individual examinee in accordance with the unknown ability level of the examinee, from the prearranged item pool, which consists of a large number of test items measuring the same ability, or abilities. Thus the search for the examinee's ability level and the search for suitable test items for him are conducted together, aiming at estimating the examinee's ability as accurately as we wish, without spending too much time and giving the examinees too many test items. The efficiency in estimating the examinee's ability, therefore, is the essential part of the adaptive testing. We can perform adaptive testing in the form of paper-and-pencil testing, but the most effective way may be the use of computers with screen terminals. Latent trait theory provides us with a strong rationale for adaptive testing, which cannot possibly be done by classical test theory.

(VII.1) Addition of New Test Items to the Item Pool

As was pointed out in Section III.4, the approaches and methods which were observed in Chapters 3, 5 and 6, for estimating the operating characteristics of the discrete item responses are most useful in developing the item pool. When we start from scratch, the first step we must take is to develop a certain number of test items which measure the ability of our interest, to confirm their dimensionality, and, selecting a suitable model, or models, to find out the operating characteristics of these test items. In so doing, we need a certain norm group of examinees to administer these core test items to obtain the basic data, and also this process includes the elimination of unfit test items, or their modifications. After this has been completed, if we wish to add more test items to our item pool, we may develop more test items and estimate their operating characteristics using one of our combinations of an approach and a method. This latter process can also be used in the situation where an item pool is already there and has been used for a long time.

An advantage of this situation in adaptive testing is that we do not need the transformation of ability θ to τ, which was described in Sections III.8 and V.3, provided that we design our procedure suitably. This will be discussed in Section VII.5.

(VII.2) Weakly Parallel Tests

Weakly parallel tests have been introduced (Samejima, 1977b) in contrast to strongly parallel tests in the context of latent trait theory. Two tests are strongly parallel if:

(1) they have the same number of items, and

(2) there is a one-to-one correspondence of each item on the first test with one and only one item on the second test with respect to the identity of the number of item score categories and the set of operating characteristics of item score categories.

In contrast to that, weakly parallel tests are any pair of tests measuring the same ability or latent trait for which the square roots of the test information functions are identical. Thus two weakly parallel tests may have:

(1) different numbers of items, and

(2) no one-to-one correspondence between the two sets of test items with respect to the number of item score categories or to the sets of operating characteristics of the item scores.

It has been pointed out (Samejima, 1977a) that in tailored testing, or computerized adaptive testing, any number of weakly parallel tests can be made by prearranging a certain amount of test information and using it as the criterion in terminating the presentation of items to individual subjects. In such procedures two different item pools are not needed, although two item pools developed for measuring the same ability or latent trait will serve just as well.

(VII.3) Use of the Amount of Test Information as the Criterion for Terminating the Presentation of New Test Items

It has been common for researchers to apply a certain degree of convergence of the current estimate of ability obtained after each test item has been presented, as the criterion for terminating the presentation of new items. This procedure, however, will result in producing different levels of accuracy of estimation at different levels of ability, or even at the same level of ability.

For the purpose of illustration, Figure 7-3-1 presents 10

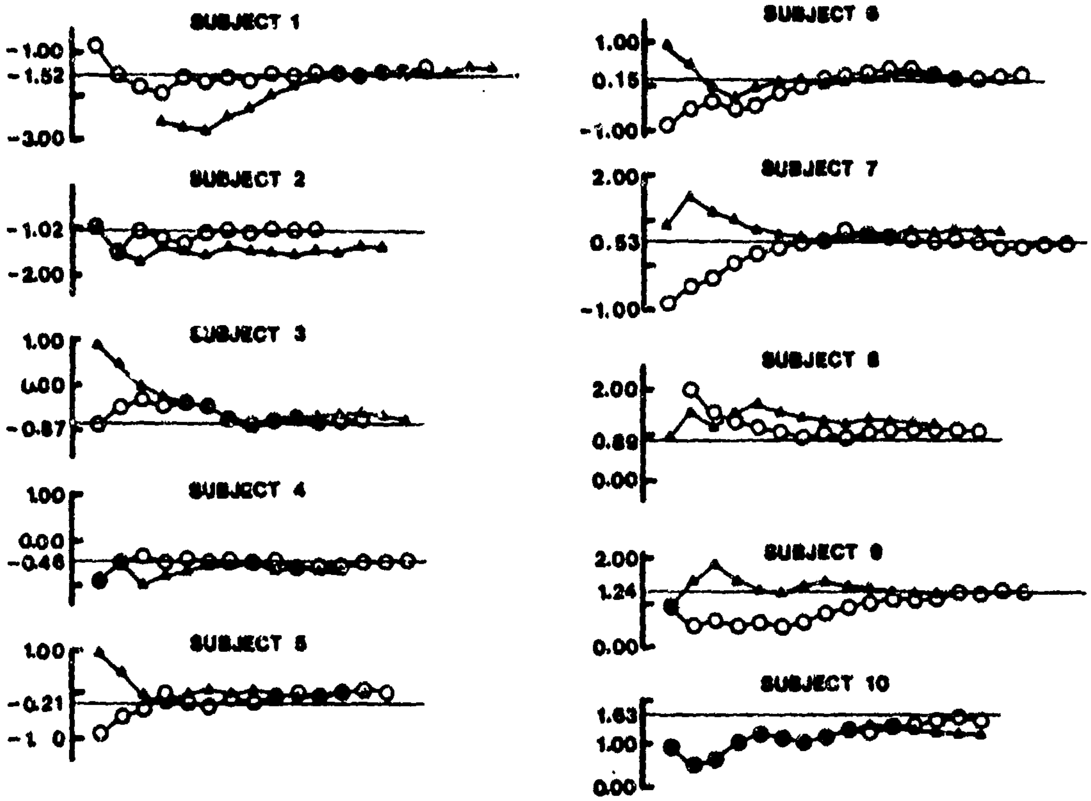


FIGURE 7-3-1

Graphic Presentation of the Change of the Local Maximum Likelihood Estimate After the Presentation of Each New Item for Each of Ten Hypothetical Examinees. Two Sessions Were Administered to Each Examinee Which Are Shown by Hollow Circles and Solid Triangles, Respectively.

graphs, each of which displays the process of convergence of the maximum likelihood estimate in a simulated tailored testing situation. The ability level of each of these 10 hypothetical examinees is shown by a number on the ordinate and the horizontal line. The item pool used for this simulation study consists of nine subsets of binary test items following the normal ogive model, whose discrimination and difficulty parameters are shown in Table 7-3-1.

TABLE 7-3-1

Item Discrimination Parameter, a_g , and Item Difficulty Parameter, b_g , of Each of the Nine Groups of Binary Test Items Used as the Item Pool in the Simulated Tailored Testing.

Item Group	a_g	b_g
1	1.20	-2.00
2	1.60	-1.50
3	2.00	-1.00
4	1.40	-0.50
5	1.80	0.00
6	1.30	0.50
7	1.70	1.00
8	1.90	1.50
9	1.50	2.00

It is assumed that each subset has a sufficiently large number of equivalent test items. There are two sessions for each examinee, which are marked with hollow circles and solid triangles in Figure 7-3-1, respectively. For each examinee in each session, binary items were selected and presented until the test information at the current value of the maximum likelihood estimate had reached 25.0 . Since the items are binary, no local maximum likelihood estimate was obtained after administration of the first item. For Subject 1, for instance, in the first session the first local maximum likelihood estimate was given after administration of the second item; and in the second session it was obtained after administration of the fifth item. It is clear from this figure that, in some cases, the current maximum likelihood estimates converged well before the test information reached 25.0 , whereas, in other cases, they had not converged yet by the time the test information reached 25.0 .

Consider, for example, Subject 4 in the first session (hollow circles) and Subject 10 in the second session (solid triangles). If the rule is made that the presentation of new items is to be terminated when the shift of the current maximum likelihood estimate is less than 0.07 twice in succession, then this will occur after the presentation of the 9th item in the former and not until the presentation of the 14th item in the latter. The corresponding values of test information are 13.370 and 23.137 , respectively. The standard error of estimation, which is the inverse of the square root of test information, is 0.273 in the former and 0.208 in the latter, i.e., approximately 76 percent of 0.273 . On the other hand, if the rule is made that the presentation of new items is to be terminated when test information has reached, say, 25.0 , at that current maximum likelihood estimate, as was the case here, the standard error of estimation would be approximately the same for all the examinees of different ability levels, i.e., 0.20 . If the estimation of each examinee's ability with the same level of accuracy is desired, there will be no doubt that the second rule is better than the first rule.

If the same level of accuracy of estimation is unnecessary, as in selection, it will be possible to prearrange a desirable test information function which is not constant for the entire range of ability in question but has a specific curve for the specific purpose. This test information function can then be used as the criterion for terminating the presentation of new items. In such a case, examinees of different levels of ability are measured with different levels of accuracy of estimation and yet the resulting selection will be conducted as accurately as is desirable if the appropriate information curve is used.

The above are only two examples of many possibilities. In any case, the use of test information functions as the criterion for terminating the presentation of new items in tailored testing permits control of the level of accuracy of estimation to serve the purposes of testing; it is impossible to do so if the convergence of the current maximum likelihood estimate is used as the criterion. The adoption of the test information function as the criterion, therefore, is strongly recommended, rejecting the convergence of the current maximum likelihood estimate, which makes the accuracy of estimation arbitrary.

(VII.4) Test Information Function and Standard Error of Estimation

One of the many advantages of latent trait theory over classical test theory is that the standard error of measurement is defined more meaningfully, as a function of the latent trait θ . It is defined as the inverse of the square root of the test information function, and is most meaningful when the test information function assumes a high enough value so that the conditional distribution of the error ε , given θ , is approximately normal. When a prearranged value of the test information function is used as the criterion for terminating the presentation of new items in adaptive testing, however, consideration must be given to the relationship between the test information function and the standard error of estimation. Figure 7-4-1 presents this relationship.

As can be seen in this figure, the latter is a strictly decreasing function of the former; yet the amount of decrement in the standard error of estimation is conspicuous for the initial increase of the test information function. It is more or less stabilized, however, after the test information function reaches 20.0 . For instance, for $I(\theta) = 6.25$ the standard error of estimation is 0.4 ; this becomes 0.2 , i.e., one-half, when $I(\theta) = 25.0$. On the other hand, to make the standard error of estimation one-fourth of 0.4 , i.e., 0.1 , the test information must be 100.0 . This suggests that, in adaptive testing, we must balance the increase in the number of test items with the decrease in the standard error of estimation, and find out a suitable criterion.

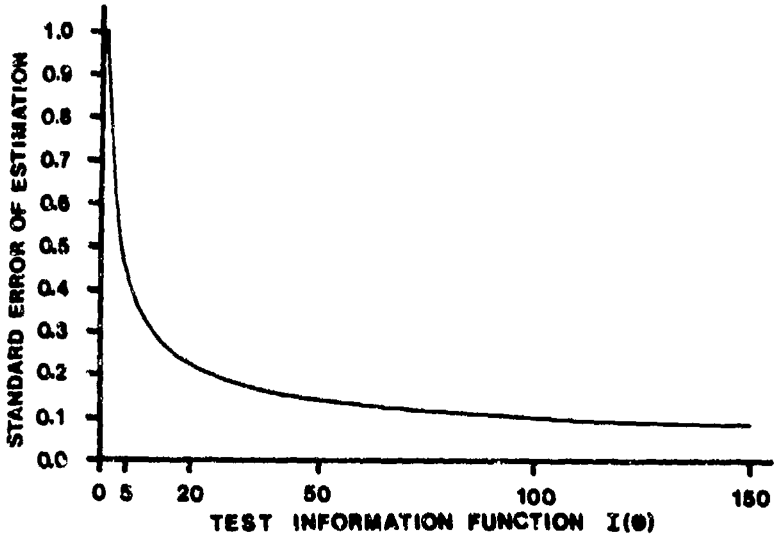


FIGURE 7-4-1

Functional Relationship between Test Information Function and Standard Error of Estimation.

(VII.5) <u>Old Test for Item Calibration</u>

It should be noted that, in adaptive testing, we can prearrange the target square root of test information, and use the function as the criteria for terminating the presentation of new items to individual examinees. This target function does not specify a single subtest from the item pool, but it provides us with a set of different, individualized subtests. If we repeat this process, we will obtain more than one such set of individualized subtests, which are weakly parallel to one another. We notice that, in spite of this difference, we may use such a set, or sets, of subtests as our Old Test, in estimating the operating characteristics of the discrete item responses to new test items, with the prearranged square root of the test information function for the interval of ability of our interest. This is a remarkable characteristic of the approaches and methods developed in the present research, when they are applied to the adaptive testing situation.

It should also be noted that, because of this characteristic, there is no need for us to transform ability θ to τ, since we can prearrange a substantially large constant value for the target square root of the test information function for our Old Test. The process of estimating the operating characteristics for new items becomes, therefore, much more simplified than the one we must use when our Old Test is a fixed test, since, under the ordinary circumstances, it is extremely difficult to develop a fixed test which has a constant amount of test information for the range of ability of our interest.

(VII.6) Adaptive Testing Using Graded Test Items

With the consideration described in earlier sections, a hypothetical tailored testing situation was constructed, using six different item pools. The first item pool consists of eleven types of graded items, each of which had four graded item score categories. Each item follows the normal ogive model, which is given by (3.6), and the three difficulty parameters, b_{x_g} for x_g=1,2,3, for each of the eleven types of graded items are presented in Table 7-6-1. The

TABLE 7-6-1

Three Difficulty Parameters for Each of the Eleven Types of Graded Test Items Which Are Common to the Three Different Item Pools.

Item	$x_g = 1$	$x_g = 2$	$x_g = 3$
1	-3.0	-2.5	-2.0
2	-2.5	-2.0	-1.5
3	-2.0	-1.5	-1.0
4	-1.5	-1.0	-0.5
5	-1.0	-0.5	0.0
6	-0.5	0.0	0.5
7	0.0	0.5	1.0
8	0.5	1.0	1.5
9	1.0	1.5	2.0
10	1.5	2.0	2.5
11	2.0	2.5	3.0

discrimination parameters, a_g , for these eleven types of items are uniformly 1.0 . The second item pool also has eleven types of graded items with the same number of item score categories and values of the difficulty parameters, but the common discrimination parameter, a_g , is 2.0 instead of 1.0 . The third item pool is the same as the first and the second, except that $a_g = 3.0$. The other 3 item pools are identical to the first set of 3 item pools, except that the items are binary items and the difficulty parameters are those shown in the column indicated as $x_g = 2$ in Table 7-6-1. It is assumed that in each item pool, there are a substantially large number of items of each type.

The criterion square root of the test information was set as $[I(\theta)]^{1/2} = 4.65$, the same constant which was used in our original Old Test. This value can also be considered as the reasonable compromise suggested in Section VII.4 . The standard error of estimation is approximately 0.215 . One hundred hypothetical subjects were used in each tailored testing situation. Their ability levels are -2.475 through 2.475 with an interval of 0.05 , i.e., the same set of one hundred ability levels as we used before (cf. Section III.3). In each pair of adaptive testing situations in which the same discrimination parameter was used, the same seed number was used to produce the same sequence of random numbers. The first item presented to every subject was item 6, which is the item with intermediate difficulty. If the subject's item score was 0 , then the easiest item, item 1, was presented repeatedly until an item score other than 0 was obtained. If the subject's score on item 6 was 4 , then the most difficult item, item 11, was presented repeatedly until an item score other than 4 was obtained. After that, the tentative maximum likelihood estimate was computed, and the computer presented an item for which the amount of test information was greatest at that value of θ . This process was repeated until the square root of the test information function at the current maximum likelihood estimate reached the criterion, 4.65 .

Tables 7-6-2 through 7-6-4 present the frequency distributions of the number of items needed for the hypothetical tailored testing for individual subjects with the criterion 4.65 in each of the two situations, for $a_g = 1.0$, 2.0 and 3.0 , respectively. A substantial difference between the two frequency distributions are observed. The mean number of items is 36.92 for the binary case and 27.98 for the graded case for $a_g = 1.0$, indicating that only 75.8 percent of the items were necessary in the graded case as compared to the binary case. These numbers are 11.97 and 7.88 for the cases where $a_g = 2.0$, and 7.38 and 4.56 where $a_g = 3.0$, and the corresponding percentages are 65.8 and 61.8 for these two pairs, respectively. This result indicates the high efficiency of the graded test items in adaptive testing, in preference to binary test items. This is especially true when we have large values for the discrimination parameters.

TABLE 7-6-2

Frequency Distribution of the Number of Items Used in Hypothetical Tailored Testing. $a_g = 1.0$.

Number of Items	Binary	Graded
27		15
28		74
29		10
30		
31		1
32		
33		
34		
35	1	
36	48	
37	31	
38	9	
39	4	
40	4	
41	2	
42	1	
Total	100	100
Mean	36.92	27.98

TABLE 7-6-3

Frequency Distribution of the Number of Items Used in Hypothetical Tailored Testing. $a_g = 2.0$.

Number of Items	Binary	Graded
7		21
8		70
9		9
10	1	
11	26	
12	54	
13	10	
14	3	
15	1	
16		
17		
18	1	
Total	96	100
Mean	11.97	7.88

TABLE 7-6-4

Frequency Distribution of the Number of Items Used in Hypothetical Tailored Testing. $a_g = 3.0$.

Number of Items	Binary	Graded
3		16
4		15
5	2	66
6	2	3
7	55	
8	36	
9	4	
Total	99	100
Mean	7.38	4.56

(VII.7) Bayesian vs. Maximum Likelihood Estimation in Adaptive Testing

As we have observed in Sections VI.1 and VI.2, the use of a prior in ability estimation provides us with biases which we may wish to avoid.

In adaptive testing, it is typical for researchers to use a normal density function as the prior. Figure 7-7-1 presents four functions, i.e., the standard normal density function, $n(0,1)$ (solid line), and three approximations to $n(0,1)$. Each of these three approximations is the product of two functions, $P_i(\theta)$ and $[1-P_j(\theta)]$, which are given by the normal ogive functions such that

$$(7.1) \qquad P_i(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a_i(\theta-b_i)} e^{-u^2/2}\,du$$

and

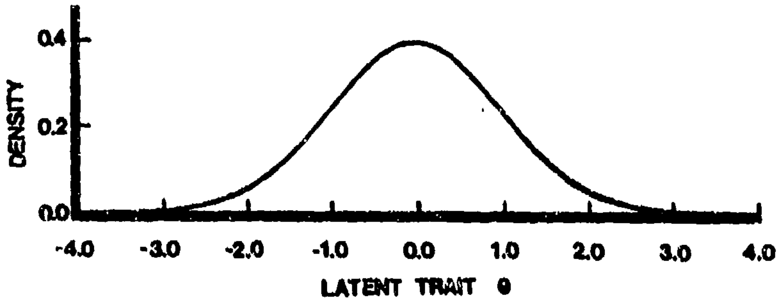


FIGURE 7-7-1

Comparison of Three Approximations with the Normal Density Function, n(0,1) (Solid Line). These Approximations Are the Products of a Normal Ogive Function and Another Subtracted From Unity, Which Equal n(0,1) at $\theta = 0.3$ (Dotted Line), $\theta = 0.6$ (Broken Line) and $\theta = 0.9$ (Dashed Line), Respectively.

$$(7.2) \qquad P_j(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a_j(\theta-b_j)} e^{-u^2/2}\, du \quad ,$$

where $a_i = a_j$ and $b_i = -b_j$. These two parameters, a_i and b_i , are 0.94810 and -0.35454 for the function drawn by a dotted line in Figure 7-7-1, 0.94980 and -0.35391 for the one drawn by a broken or long, dashed line, and 0.95259 and -0.35287 for the one drawn by a short, dashed line, respectively. These three approximations are obtained by setting the product of the two functions equal to the standard normal density function at $\theta = 0.3$, $\theta = 0.6$ and $\theta = 0.9$, respectively, in addition to $\theta = 0.0$. We notice that these four curves, including n(0,1) , in Figure 7-7-1 are practically indistinguishable.

We notice that the formulas in (7.1) and (7.2) are identical with the item characteristic function in the normal ogive model. This implies that the prior, n(0,1) , is practically the same as the product of the two operating characteristics of the hypothetical binary items, i and j , for the response pattern, (1,0) . The Bayes modal estimator with the prior n(0,1) can be considered, therefore, as the maximum likelihood estimator, obtained from the response pattern V plus two additional responses, 1 and 0 , to the hypothetical binary items, i and j . Note that these two additional item responses are always 1 and 0 , regardless of the true ability level.

In order to observe how the prior affects the resultant ability estimation in adaptive testing, a simulation study was conducted (RR-80-3) by using the hypothetical item pool, which was described in Section VII.3 . We assume eleven hypothetical examinees, whose ability levels are -2.25 , -1.75 , -1.25 , -0.75 , -0.25 , 0.00 , 0.50 , 1.00 , 1.50 , 2.00 and 2.50 , respectively. We also assume four different situations, in one of which the maximum likelihood estimation is applied for the ability estimation, and in the other three Bayes modal estimation is used, with three different priors, n(0.0,1.0) , n(0.0,0.8) and n(0.0,0.5) , respectively. In the first situation of maximum likelihood estimation, an item from group 5 is always chosen as the first item to present to an examinee, and, depending upon the examinee's response to this item, the second item is chosen either from group 1 or group 9. That is to say, if the examinee's response to the first item is correct, then the second item is chosen from group 9, i.e., the most difficult item group, and, if it is incorrect, then the second item is chosen from group 1, the easiest item group. The examinee will stay with the same item group for the following items, until he fails in answering an item correctly if it is group 9, and until he succeeds in answering an item correctly if it is group 1. Thereafter, since every current likelihood function has a local maximum, an item from the item group whose item information function, $I_g(\theta)$, which is defined by (3.9) , is the greatest at the value of current maximum likelihood estimate is chosen and presented next, and this will go on until the amount of test information at the current maximum likelihood estimate reaches or exceeds a certain criterion. All the responses of the hypothetical examinees are calibrated by the Monte Carlo method.

In Bayesian estimation, the first estimate is the modal point of the prior. The second item is an item chosen from the item group whose item information function, $I_g(\theta)$, is the greatest at the modal point of the prior, and the third item is from the item group whose item information function is the greatest at the current Bayes modal estimate, and so forth, and the presentation of a new item is terminated when the amount of test information at the current estimate of the examinee's ability has reached the same criterion used in the maximum likelihood estimation.

Figure 7-7-2 presents the results of these two ability estimations, which were obtained by using the prior, n(0.0,0.8) , and without using any priors, by solid circles and solid triangles, respectively. In this figure, $\theta = 0.0$, and the prior did not interfere with the convergence of the ability estimate. In contrast to this result, Figure 7-7-3 presents another case in which $\theta = -2.25$. In this figure, substantial differences between the two processes of the maximum likelihood estimation and the Bayes modal estimation are observed, in the latter of which the convergence is much slower, fighting off the effect of the prior. These two examples typically illustrate the bias caused by the prior.

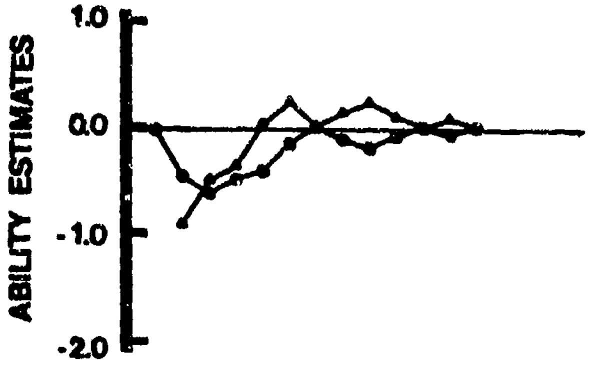


FIGURE 7-7-2

Successive Maximum Likelihood Estimates (Triangles) and Bayes Modal Estimates (Circles) in the Simulated Tailored Testing with n(0.0,0.8) as the Prior for a Hypothetical Examinee Whose Ability Level is 0.00 .

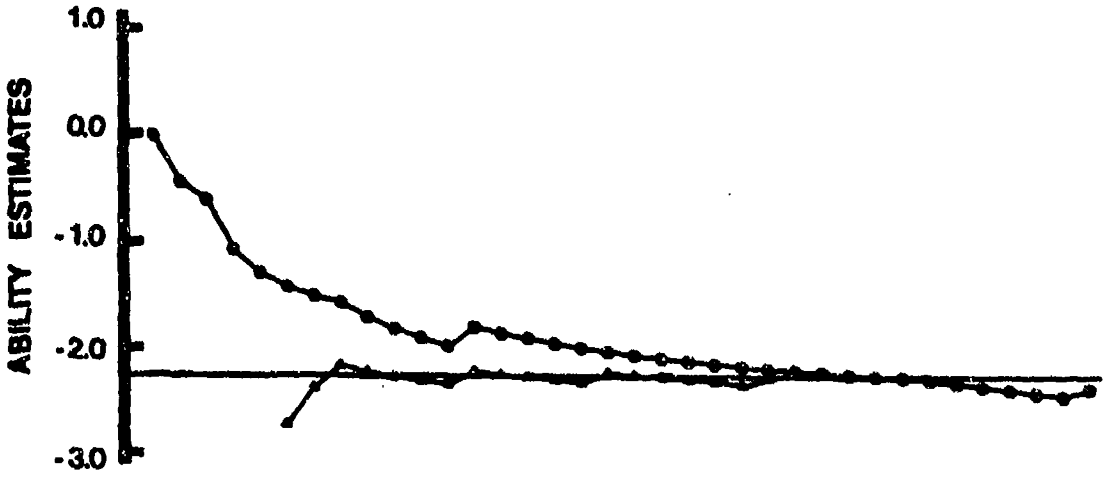


FIGURE 7-7-3

Successive Maximum Likelihood Estimates (Triangles) and Bayes Modal Estimates (Circles) in the Simulated Tailored Testing with n(0.0,0.8) as the Prior for a Hypothetical Examinee Whose Ability Level is -2.25 .

REFERENCES

[1] Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977(a), 1, 233-247.

[2] Samejima, F. Weakly parallel tests in latent trait theory with some criticisms on classical test theory. Psychometrika, 1977(b), 42, 193-198.

VIII Constant Information Model

Researchers get interested in finding out what kind of test item provides us with a larger amount of information in comparison with others. They seldom pay attention, however, to the fact that there exists some constancy in the amount of item information. In this chapter, we shall observe such aspects of information functions, introduce a new model for the binary test item, which is called Constant Information Model, and discuss its practical implications and usefulness in the estimation of the operating characteristics of discrete item responses.

(VIII.1) Constancy of Information under the Transformation of the Latent Trait

Let τ be any strictly increasing transformation of ability θ . The relationships between the two sets of information functions, i.e., $I_{x_g}(\theta)$, $I_g(\theta)$, $I_V(\theta)$ and $I(\theta)$ versus $I^*_{x_g}(\tau)$, $I^*_g(\tau)$, $I^*_V(\tau)$ and $I^*(\tau)$, have been given in Section III.8 , while the original definitions of the first set of information functions are given in Section III.4 . It should be noted that the area under the curve of the item information function, and that of the test information function, do change with the transformation of ability θ to τ , since there are such relationships that

$$(8.1) \qquad \int_{\underline{\tau}}^{\bar{\tau}} I^*_g(\tau)\, d\tau = \int_{\underline{\theta}}^{\bar{\theta}} I_g(\theta) \frac{d\theta}{d\tau}\, d\theta \quad ,$$

and

$$(8.2) \qquad \int_{\underline{\tau}}^{\bar{\tau}} I^*(\tau)\, d\tau = \int_{\underline{\theta}}^{\bar{\theta}} I(\theta) \frac{d\theta}{d\tau}\, d\tau \quad ,$$

where $\underline{\theta}$ and $\bar{\theta}$ are the lower and upper endpoints of the range of θ and

$$(8.3)\qquad \begin{cases} \underline{\tau} = \tau(\underline{\theta}) \\ \bar{\tau} = \tau(\bar{\theta}) \end{cases}$$

are those of the range of the transformed variable τ .

If we consider the integration of the square root of each information function, however, we obtain

$$(8.4)\qquad \int_{\underline{\tau}}^{\bar{\tau}} [I_g^*(\tau)]^{1/2}\, d\tau = \int_{\underline{\theta}}^{\bar{\theta}} [I_g(\theta)]^{1/2}\, d\theta \quad ,$$

and

$$(8.5)\qquad \int_{\underline{\tau}}^{\bar{\tau}} [I^*(\tau)]^{1/2}\, d\tau = \int_{\underline{\theta}}^{\bar{\theta}} [I(\theta)]^{1/2}\, d\theta \quad .$$

Thus the area under the curve of the square root of the item information function, and that of the test information function, are unchanged throughout the transformation of the latent trait by any strictly increasing function, $\tau(\theta)$.

We recall that ability θ was transformed to τ by the polynomial given by (5.13) when we used one of the nine subtests of the original Old Test, i.e., Subtests 1 through 9, as our Old Test (cf. Section VI). The above fact implies that, in so doing, the totality of the square root of the test information function of our Old Test was kept constant.

(VIII.2) Constancy of Item Information for a Specified Model

The finding in the preceding section can be generalized further to the constancy of the square root of the item information function for items which follow the same model, as long as the set of operating characteristics for an arbitrarily selected test item which belongs to the model can produce one for any other test item which follows the same model. To give an example, suppose that item g has an item characteristic function in the normal ogive model, such that

$$(8.6) \qquad P_g(\theta) = [2\pi]^{-1/2} \int_{-\infty}^{a_g(\theta-b_g)} \exp[-t^2/2]\, dt \quad ,$$

where a_g (>0) and b_g are the item discrimination and difficulty parameters, respectively. Suppose that we wish to transform ability θ to τ by the linear transformation such that

$$(8.7) \qquad \tau = a_g(\theta-b_g)\ a_g^{*-1} + b_g^* \quad ,$$

where a_g^* is an arbitrary positive constant and b_g^* is any constant. We can write for the item characteristic function, $P_g^*(\tau)$, of item g resulting from the transformation of θ to τ

$$(8.8) \qquad P_g^*(\tau) = [2\pi]^{-1/2} \int_{-\infty}^{a_g^*(\tau-b_g^*)} \exp[-t^2/2]\, dt \quad .$$

It is obvious that $P_g^*(\tau)$ thus obtained belongs to the normal ogive model. From the finding obtained in the preceding section, therefore, the constancy holds for the totality of the square root of the item information function over the transformation of θ to τ . Note that this is true for any arbitrarily chosen values for a_g^* and b_g^* , as long as a_g^* is positive. Let h be any other binary test item which also follows the normal ogive model. We can write

$$(8.9) \qquad P_h(\theta) = [2\pi]^{-1/2} \int_{-\infty}^{a_h(\theta-b_h)} \exp[-t^2/2]\, dt \quad .$$

If we set $a_g^* = a_h$ and $b_g^* = b_h$, then (8.8) provides us with an identical curve with that of (8.9) . The area under the square root of the item information function, $[I_g^*(\tau)]^{1/2}$, therefore, will equal that of $[I_h(\theta)]^{1/2}$. The constancy of item information holds over any binary test items which belong to the same model, i.e., the normal ogive model.

For the purpose of illustration, Figures 8-2-1 and 8-2-2 present the item information functions and their square roots for three items, all of which belong to the normal ogive model with

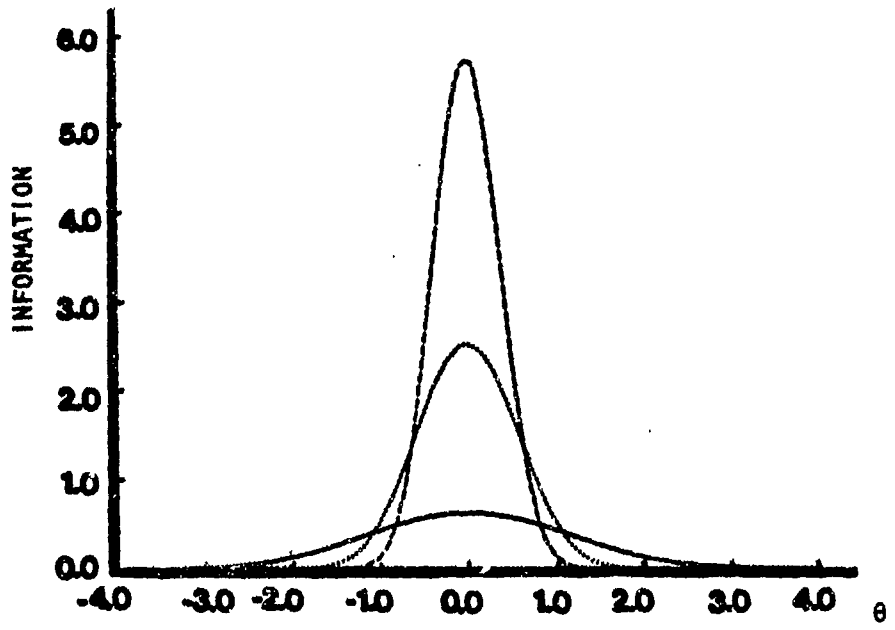


FIGURE 8-2-1

Item Information Functions of Three Binary Items, Which Follow the Normal Ogive Model, with the Common Difficulty Parameter, $b_1 = b_2 = b_3 = 0.0$, and the Discrimination Parameters, $a_1 = 1.0$ (Solid Curve), $a_2 = 2.0$ (Dotted Curve) and $a_3 = 3.0$ (Dashed Curve), Respectively.

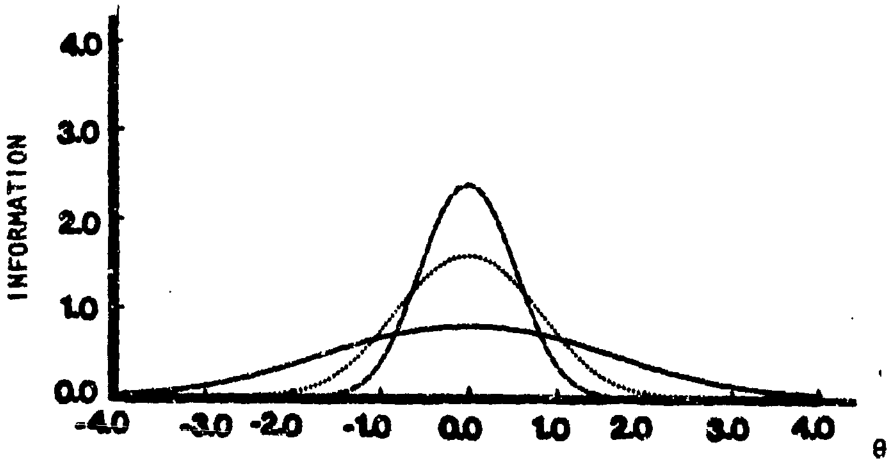


FIGURE 8-2-2

Square Roots of the Item Information Functions of Three Binary Items, Which Follow the Normal Ogive Model, with the Common Difficulty Parameter, $b_1 = b_2 = b_3 = 0.0$, and the Discrimination Parameters, $a_1 = 1.0$ (Solid Curve), $a_2 = 2.0$ (Dotted Curve) and $a_3 = 3.0$ (Dashed Curve), Respectively.

$a_1 = 1.0$, $a_2 = 2.0$ and $a_3 = 3.0$, and $b_1 = b_2 = b_3 = 0.0$, respectively. We can see that in Figure 8-2-1 the three areas are substantially different from one another, while those in Figure 8-8-2 are equal.

(VIII.3) Constancy of Item Information for a Set of Models

In this section, we only consider binary test items. Consider a set of test items which follow different models, but whose item characteristic functions are strictly increasing in θ , and satisfy

$$(8.10) \quad \begin{cases} \lim_{\theta \to \underline{\theta}} P_g(\theta) = 0 \\ \lim_{\theta \to \overline{\theta}} P_g(\theta) = 1 \quad . \end{cases}$$

Let h denote another, arbitrarily chosen test item which follows a different model, which satisfies the above two conditions. The transformation of θ to τ in such a way that

$$(8.11) \quad \tau = P_h^{-1}[P_g(\theta)]$$

provides us with the item characteristic function, $P^*_g(\tau)$, for item g with respect to the transformed latent trait τ , which is identical with $P_h(\theta)$. The constancy of item information holds, therefore, for item g and item h on the ability scale θ , in spite of the fact that they belong to different models.

Figure 8-3-1 illustrates the square roots of the item information functions of three binary test items, g , h and j , which follow the normal ogive model, the logistic model and the linear model, respectively. The item characteristic functions of item h and j are given as follows.

$$(8.12) \quad P_h(\theta) = [1 + \exp\{-Da_h(\theta-b_h)\}]^{-1} \quad -\infty < \theta < \infty \quad .$$

$$(8.13) \quad P_j(\theta) = (\theta-\alpha_j)(\beta_j-\alpha_j)^{-1} \quad \alpha_j < \theta < \beta_j \quad .$$

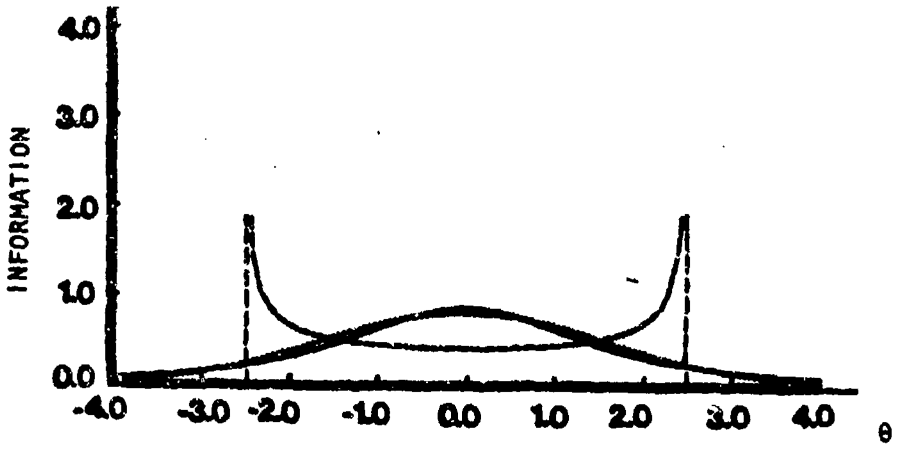


FIGURE 8-3-1

Square Roots of the Item Information Functions of Items g, h and j, Which Follow the Normal Ogive Model with $a_g = 1.0$ and $b_g = 0.0$ (Dotted Curve), the Logistic Model with $D = 1.7$, $a_h = 1.0$ and $b_h = 0.0$ (Solid Curve) and the Linear Model with $\alpha_j = -2.5$ and $\beta_j = 2.5$ (Dashed Curve).

The reader is directed to Chapter 3 of RR-79-1 for the relationships among these three models.

It should be noted that the same principle holds for any other sets of models, each of which has common characteristics of its own, as the present set of models has the strictly increasing property in item characteristic functions and the satisfaction of (8.10) . It will be improper, however, to consider a set of models for which the item information function is meaningless, like the type of the three-parameter logistic or normal ogive models, for the reason the author has pointed out (Samejima, 1973).

(VIII.4) Exact Area under the Square Root of the Item Information Function

We notice that the common area under the square root of the item information function for all the binary test items, whose item characteristic functions are strictly increasing in θ and satisfy (8.10) , can be obtained by integrating $[I_g(\theta)]^{1/2}$ for any arbitrarily chosen item g . This area equals π , or approximately 3.14159 . The following process is an example, in which the logistic model has been chosen.

(8.14) $$\int_{-\infty}^{\infty}[I_h(\theta)]^{1/2}\,d\theta = Da_h\int_{-\infty}^{\infty}[\exp\{Da_h(\theta-b_h)\}]^{1/2}\,[1+\exp\{Da_h(\theta-b_h)\}]^{-1}\,d\theta \quad .$$

(8.15) $$\theta^* = [\exp\{Da(\theta-b_h)\}]^{1/2} \quad .$$

(8.16) $$\frac{d\theta}{d\theta^*} = 2\,(Da_h)^{-1}\,\theta^{*-1} \quad .$$

(8.17) $$\int_{-\infty}^{\infty}[I_h(\theta)]^{1/2}\,d\theta = Da_h\int_{0}^{\infty}\theta^*(1+\theta^{*2})\;2(Da_h)^{-1}\theta^{*-1}\,d\theta^* = 2\int_{0}^{\infty}(1+\theta^{*2})^{-1}\,d\theta^* = 2\tan^{-1}\theta^*\Big|_{0}^{\infty} = \pi \quad .$$

It will be just as easy to demonstrate it if we choose the linear model instead of the logistic model (cf. Chapter 4, RR-79-1).

(VIII.5) Constant Information Model

To represent the type of models which satisfy the two conditions described in Section VIII.3 , we shall consider a model which provides us with a constant value for the square root of the item information function for the interval of θ , $[\underline{\theta},\bar{\theta}]$. Let g denote such a binary test item. It is obvious that the interval, $[\underline{\theta},\bar{\theta}]$, is a finite interval, since the area of the rectangle given by this interval and the constant square root of the item information function, C , is a finite value, i.e., π . Thus we can write

(8.18) $$\bar{\theta} - \underline{\theta} = \pi C^{-1} \quad .$$

Thus the length of the interval of θ depends upon the constant item information C .

We find that the model described by

(8.19) $$P_g(\theta) = \sin^2[a_g(\theta-b_g) + (\pi/4)]$$

is the one we have looked for, if we set the parameter a_g such that

$$(8.20) \qquad a_g = C/2 \quad ,$$

with the range of θ such that

$$(8.21) \qquad [-\pi a_g^{-1}/4] + b_g \leq \theta \leq [\pi a_g^{-1}/4] + b_g$$

Since we have

$$(8.22) \qquad Q_g(\theta) = 1 - P_g(\theta) = \cos^2[a_g(\theta-b_g) + (\pi/4)] \quad ,$$

and

$$(8.23) \qquad \frac{\partial}{\partial\theta} P_g(\theta) = 2 \sin [a_g(\theta-b_g) + (\pi/4)] \cdot \cos [a_g(\theta-b_g) + (\pi/4)] \cdot a_g = 2\, a_g\, [P_g(\theta)Q_g(\theta)]^{1/2} = C\, [P_g(\theta)Q_g(\theta)]^{1/2} \quad ,$$

we obtain

$$(8.24) \qquad I_g(\theta) = [\frac{\partial}{\partial\theta} P_g(\theta)]^2 [P_g(\theta)Q_g(\theta)]^{-1} = C^2 \quad .$$

We can see from (8.19) that this model provides us with point symmetric item characteristic functions with $(b_g, 0.5)$ as the point of symmetry, just like the normal ogive model, the logistic model and the linear model. The parameter b_g can be called, therefore, difficulty parameter, just as in the normal ogive and logistic models. It is obvious from (8.23) that the parameter a_g is proportional to) the slope of the line tangent to $P_g(\theta)$ at $\theta = b_g$, just as in these two models, so it can be called discrimination parameter. The meaning of this parameter is more obvious in (8.20), i.e., the fact that the amount of item information solely depends upon the parameter a_g .

We shall call this model, which is presented by (8.19), the Constant Information Model. This model has an important role in the estimation of the operating characteristics of item response categories, which will be described in the following section.

Figure 8-5-1 presents a few examples of the item characteristic

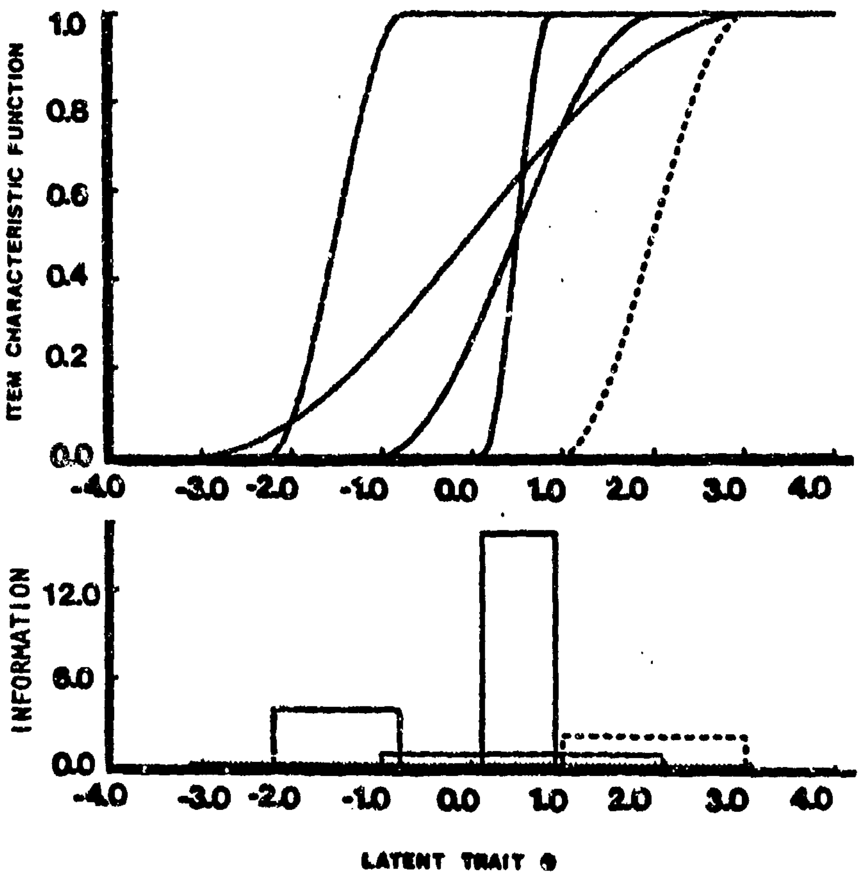


FIGURE 8-5-1

Item Characteristic Functions (Upper Graph) and the Item Information Functions (Lower Graph) of Five Binary Items Following the Constant Information Model. The Item Parameters Are: $a_1 = 0.25$ and $b_1 = 0.00$ (Smaller Dots), $a_2 = 0.50$ and $b_2 = 0.50$ (Shorter Dashes), $a_3 = 0.75$ and $b_3 = 2.00$ (Larger Dots), $a_4 = 1.0$ and $b_4 = -1.5$ (Longer Dashes), and $a_5 = 2.00$ and $b_5 = 0.50$ (Solid Line).

function of the Constant Information Model, together with the corresponding item information functions.

The item response information function, $I_{x_g}(\theta)$, in the Constant Information Model can be written as

$$(8.25) \qquad I_{x_g}(\theta) \begin{cases} = 2a_g^{\;2} \sec^2[a_g(\theta-b_g)+(\pi/4)] = 2a_g^{\;2}[Q_g(\theta)]^{-1} > 0 & \text{for } x_g = 0 \\ = 2a_g^{\;2} \csc^2[a_g(\theta-b_g)+(\pi/4)] = 2a_g^{\;2}[P_g(\theta)]^{-1} > 0 & \text{for } x_g = 1 \end{cases} .$$

Figure 8-5-2 illustrates these two item response information functions for an item with the parameters, $a_g = 0.25$ and $b_g = 0.00$, together with the constant item information function (= 0.25) . From (3.12)

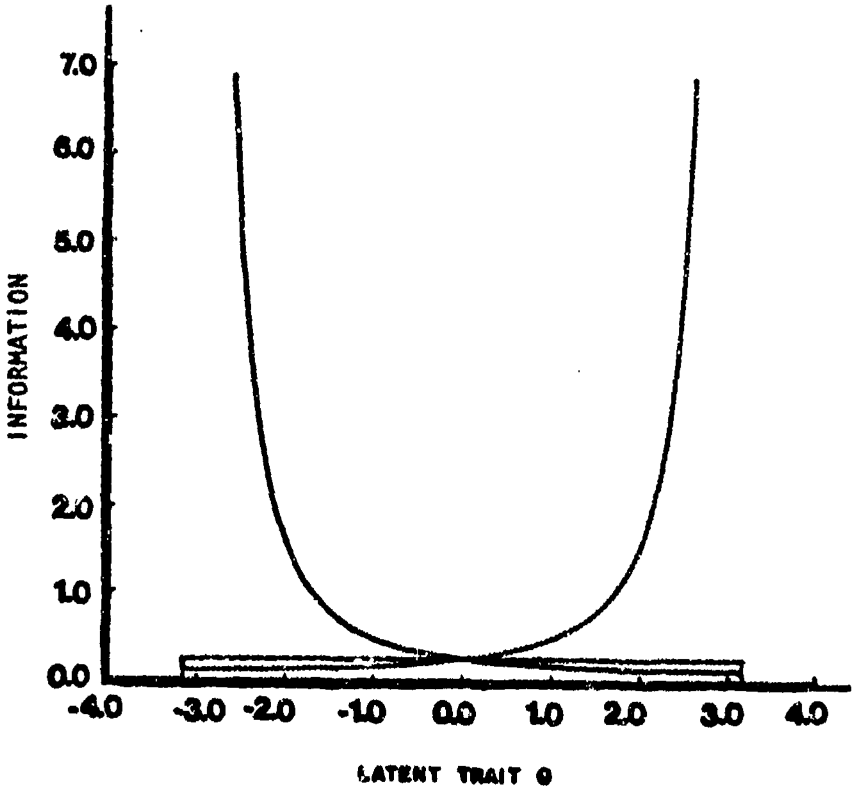


FIGURE 8-5-2

Item Response Information Functions of an Item Following the Constant Information Model, with the Parameters, $a_g = 0.25$ and $b_g = 0.00$, for $x_g = 0$ (Dotted Curve) and for $x_g = 1$ (Solid Curve), Together with the Constant Item Information Function (Dashed Curve).

and (8.25) we can write for the response pattern information function

$$(8.26) \qquad I_V(\theta) = 2 \sum_{x_g \epsilon V} a_g{}^2[P_g(\theta)]^{-x_g}[Q_g(\theta)]^{x_g-1} \quad ,$$

and, finally, the test information function is given by

$$(8.27) \qquad I(\theta) = 4 \sum_{g=1}^{n} a_g{}^2 \quad .$$

(VIII.6) Use of Constant Information Model for a Set of Equivalent Test Items Which Substitutes for the Old Test

In our combinations of a method and an approach, we need Old Test, or a set of test items whose operating charactersitics are known (cf. Chapter 3). In some situations, however, we may lift this restriction, with the effective use of Constant Information Model.

Suppose, for developing the new item pool, a substantial number of test items are administered to a substantial number of examinees, and there exists a subset of equivalent binary items among these items. In this situation, we can use this subset of items as the substitute for the Old Test.

It has been shown by Birnbaum (Birnbaum, 1968) that, when the test consists of n equivalent, binary items, the simple test score t, which is the sum total of the n binary item scores, is a minimal sufficient statistic for the response pattern V. In such a case, we have

$$(8.28) \qquad t = nP_g(\theta) \quad ,$$

and the maximum likelihood estimate $\hat{\theta}$ is given by

$$(8.29) \qquad \hat{\theta} = P_g^{-1}(t/n) \quad .$$

When this common item characteristic function follows the

Constant Information Model, we obtain from (8.19) and (8.29)

$$\hat{\theta} = a_g^{-1} [\sin^{-1}(t/n)^{1/2} - (\pi/4)] + b_g \quad . \tag{8.30}$$

It is obvious from (8.21) and (8.30) that the range of $\hat{\theta}$ is given by

$$[-\pi a_g^{-1}/4] + b_g \leq \hat{\theta} \leq [\pi a_g^{-1}/4] + b_g \quad . \tag{8.31}$$

We assume that these equivalent items have a strictly increasing item characteristic function with 0 and 1 as its two asymptotes. As we have seen in previous sections, we can adjust the latent trait scale in such a way that the resulting common item characteristic function for these equivalent items follow the Constant Information Model, which is given by (8.19) . Then the response pattern of each examinee with respect to the subset of equivalent binary items is specified, and is summarized in the form of test score. The origin and unit of the latent trait are set more or less arbitrarily, say, $a_g = 0.25$ and $b_g = 0.00$. From the test score of the subset of equivalent binary items, the maximum likelihood estimate of the examinee's ability is obtained through (8.30) . The resulting set of the maximum likelihood estimates for all the examinees can be used in the same way as we use the set of maximum likelihood estimates obtained from the results of the Old Test. The operating characteristics of each of the other items can be estimated in the same way as we do when we use the Old Test. After this has been done, we can transform the latent trait in whatever way we wish.

(VIII.7) How to Detect a Subset of Equivalent Binary Items

A natural question is how to detect a subset of equivalent binary items out of the tentative item pool. In empirical sciences, it is often difficult to obtain a sufficient evidence. The second best way will be, therefore, to formulate a set of necessary evidences, and to check our data with respect to each criterion. If we find out that our data satisfy all the necessary conditions thus formulated, then we can assume that we have obtained what we wanted, until another necessary criterion becomes available and our data fail to satisfy it.

In our situation, first of all, it is necessary, though not sufficient, that the proportions correct should be the same value for all the equivalent binary items, with the allowance of sampling fluctuations. This can be checked easily, and we can find out a group of binary items which satisfy this condition, if there is any. It is also necessary that the 2 x 2 contingency tables of the bivariate frequency distributions should be symmetric and identical among all the pairs of equivalent binary items, within the allowance of sampling fluctuations. This can be checked for every pair of binary items which have passed the first selection, and, possibly, some items have to be dropped. We can go ahead to the 2^3 contingency tables after this step, to the 2^4 contingency tables, etc., if we wish.

Unlike the common belief in high discrimination power, it is desirable that these equivalent items have a low common discrimination, in addition to being substantial in number. A necessary condition for this is that the two frequencies for the response patterns (0,1) and (1,0) , which are, theoretically, the same value if the two items are equivalent, should be large, or compatible to the other two. This can be checked, therefore, in the same process for checking the equivalency of the binary items. Table 8-7-1 illustrates two typical 2 x 2

Low Discrimination Parameter

Item h Item g	$x_h = 0$	$x_h = 1$	Total
$x_g = 0$	110	243	353
$x_g = 1$	248	399	647
Total	358	642	1000

High Discrimination Parameter

Item h Item g	$x_h = 0$	$x_h = 1$	Total
$x_g = 0$	300	53	353
$x_g = 1$	58	589	647
Total	358	642	1000

TABLE 8-7-1

Two Typical 2 x 2 Contingency Tables for a Pair of Equivalent Items with a Common Low Discrimination Parameter, and for Those with a Common High Discrimination Parameter, Respectively contingency tables, one of which is for a pair of equivalent binary items which have a common low discrimination parameter, and the other is for a pair of those which have a common high discrimination parameter.

(VIII.8) Convergence of the Conditional Distribution of the Maximum Likelihood Estimate to the Asymptotic Normality When a Test Consists of Equivalent Items

In using the generalized method, we should be aware of a few problems. First of all, the constant test information provided by the subset of equivalent binary items following Constant Information Model should be substantially large, so that the normal approximation for the conditional distribution of $\hat{\theta}$, given θ , should be acceptable. On the other hand, we need a substantially wide range of ability θ for which the test information is constant, in order to make the estimation of the operating characteristics of the other items meaningful. These two are opposing factors, as is obvious from (8.20) and (8.21) . The solution for this problem is to use a substantially large number of equivalent binary items, whose common discrimination parameter is low, as was mentioned in the preceding section.

Another problem is the effect of the range of $\hat{\theta}$ on the speed of convergence of the conditional distribution of $\hat{\theta}$, given θ , to the normal distribution, $N(\theta, (n^{-1/2}C^{-1}))$. Since the range of $\hat{\theta}$ is a finite interval which is given by (8.31) , it should be expected that the truncation of the conditional distribution makes the convergence slow around the values of θ close to $(-\pi a_g^{-1}/4)+b_g$ and $(\pi a_g^{-1}/4)+b_g$, as is illustrated in Figure 8-8-1. A solution for this problem is again to use a set of equivalent binary items whose common discrimination parameter is low, so that the range of θ is wide enough to include all the examinees far inside of the two endpoints of the interval of θ . An alternative for the above solution is to exclude examinees whose $\hat{\theta}$'s are close to $(-\pi a_g^{-1}/4)+b_g$ or $(\pi a_g^{-1}/4)+b_g$. In the second solution, however, the number of examinees will be decreased and this may affect the accuracy of the estimation of the operating characteristics. It is worth noting that the solution for the first problem, is also the solution for the second problem.

If there exist more than one subset of equivalant binary items within the tentative item pool, we can make a full use of all the subsets. We follow the process described earlier for each subset of equivalent binary items, and the resultant estimated operating characteristics can be equated by appropriate transformations of the separately defined latent traits, using, say, the least squares principle, to integrate all of them into one scale.

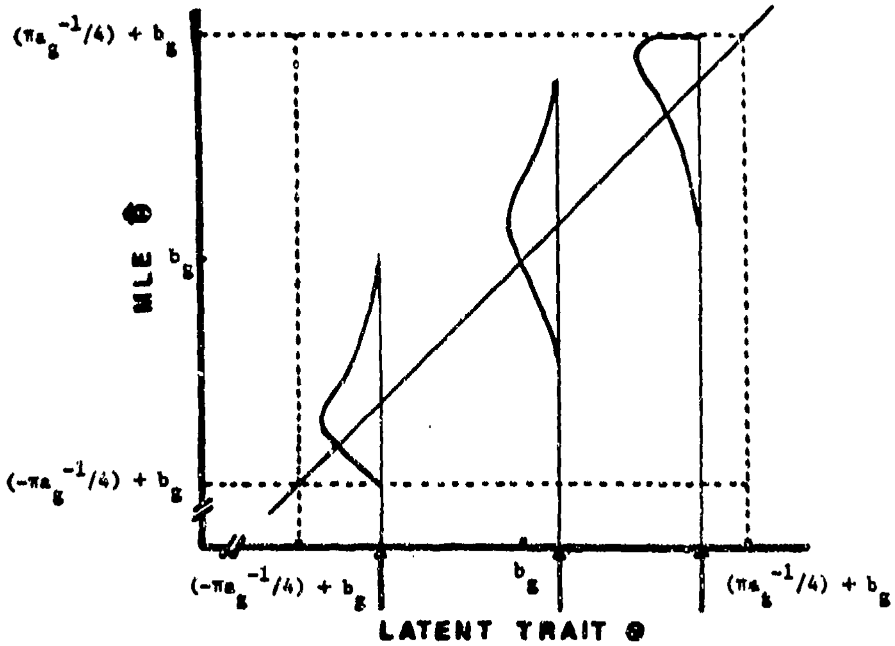


FIGURE 8-8-1

Graphical Illustration of the Conditional Density Functions of the Maximum Likelihood Estimate $\hat{\theta}$, Given the Latent Trait θ.

In order to pursue the process of convergence of the conditional distribution of the maximum likelihood estimate, givan ability, to the asymptotic normality when a test consists of n equivalent, binary test items, a Monte Carlo study was conducted (cf. RR-79-3).

For the common item characteristic function of the hypothetical equivalent, binary items, Constant Information Model with the parameters,

$$(8.32)\qquad \begin{cases} a_g = 0.25 \\ b_g = 0.00 \quad , \end{cases}$$

was used. The interval of θ for which the item information function assumes a positive constant is given by

$$(8.33)\qquad -\pi < \theta < \pi \quad ,$$

and we have for the amount of item information

$$(8.34)\qquad I_g(\theta) = 0.25 \quad .$$

As the fixed levels of the latent trait θ , eight positions were selected, i.e., -3.0 , -2.2 , -1.4 , -0.6 , 0.2 , 1.0 , 1.8 and 2.6 . A group of one hundred hypothetical examinees were assigned to each of the eight levels of ability θ , to make the total number of hypothetical examinees eight hundred. There were twenty hypothetical sessions of testing, and in each session ten equivalent, binary items were administered. An item score x_g (= 0 or 1) was calibrated by the Monte Carlo method following the Constant Information Model. After the completion of each session, the cumulative test score t was computed for each of the eight hundred hypothetical examinees. Thus after the completion of the k-th session the full test score is $10 \times k$. The maximum likelihood estimate $\hat{\theta}$ was obtained by

$$(8.35)\qquad \begin{aligned} \hat{\theta} &= P_g^{-1}[t/(10k)] \\ &= 4\sin^{-1}\{[t/(10k)]^{1/2}\} - \pi \end{aligned}$$

for each hypothetical subject, after the completion of the k-th session. As an example of slow convergence, Figure 8-8-2 illustrates

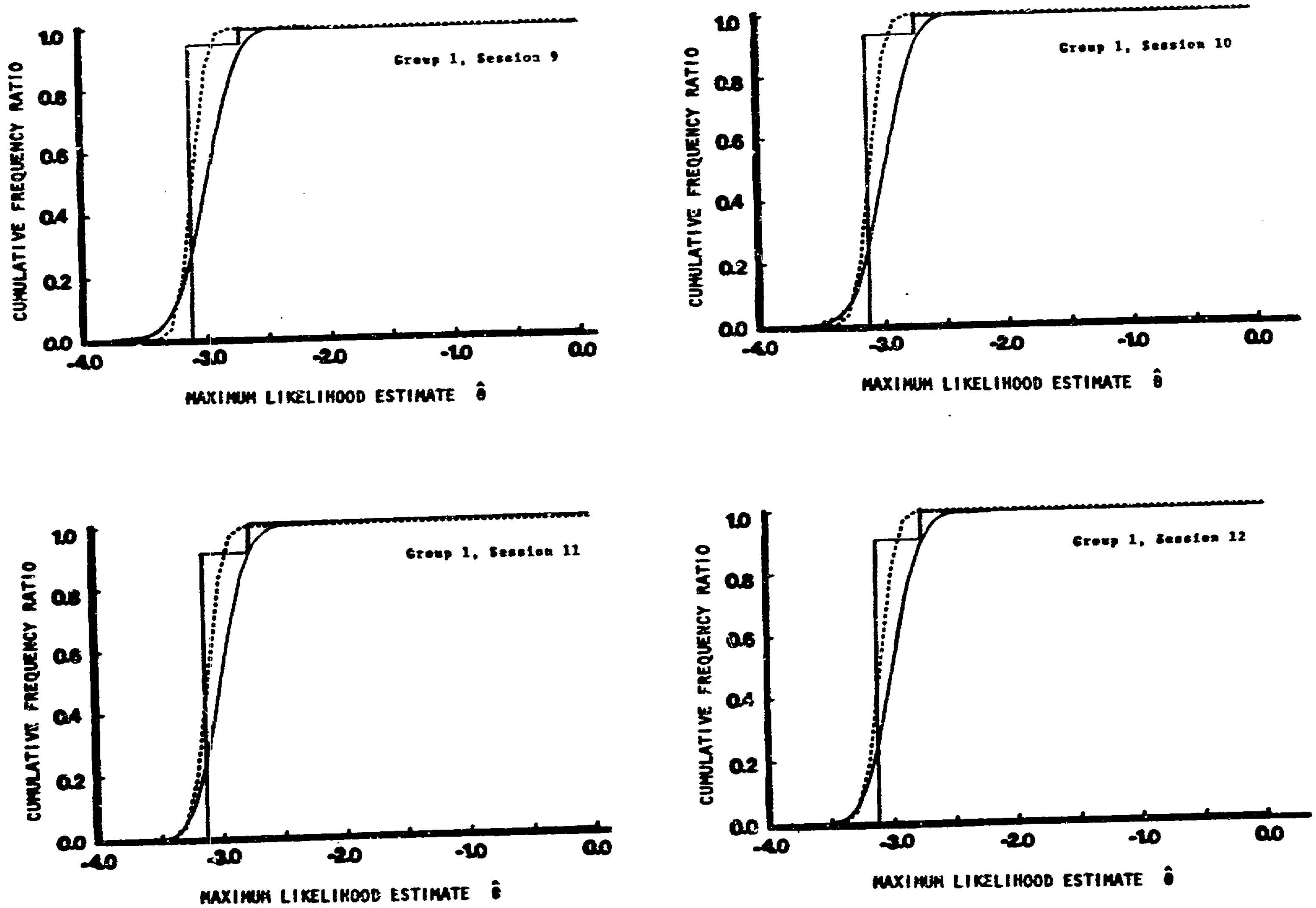


FIGURE 8-8-2

Cumulative Frequency Ratio of the Maximum Likelihood Estimate $\hat{\theta}$ of the 100 Hypothetical Examinees (Step Function) Whose Ability Levels Are Uniformly -3.0 , with the Asymptotic Normal Distribution Function (Solid Curve) and the Normal Distribution Function with the Sample Mean and Standard Deviation of $\hat{\theta}$ As Its Two Parameters (Dotted Curve), After Completing Sessions 9, 10, 11, 12, 17, 18, 19 and 20 , Respectively.

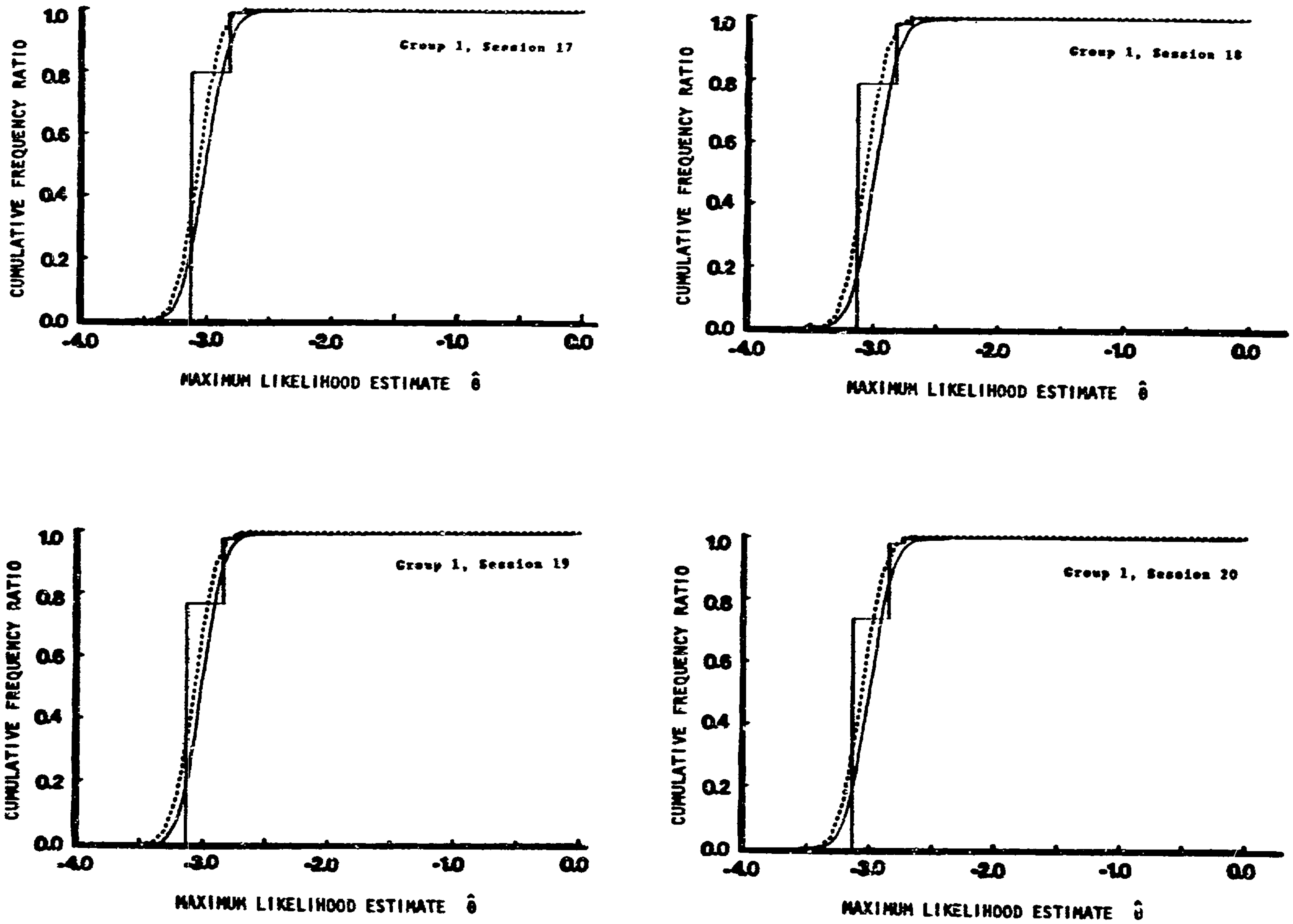


FIGURE 8-8-2 (Continued)

the resultant cumulative frequency ratios of the maximum likelihood estimates of the one hundred hypothetical examinees of group 1 by step functions, along with the normal distribution functions, $N(\theta,\{I(\theta)\}^{-1/2})$, which are drawn by solid curves, after the completions of Sessions 9, 10, 11, 12, 17, 18, 19 and 20 , respectively. In the same figure, also presented are the corresponding normal distribution functions with the sample mean and standard deviation of $\hat{\theta}$ as the two parameters, by dotted curves. We can see in this figure that the two normal distribution functions are still distinctly apart, even after all the twenty sessions.

Figure 8-8-3 presents the corresponding set of results for Group 5 , as an example of fast convergence. We can see in this figure that the approximation is good enough even after Session 9 . For the details of this study, the reader is directed to the research report, RR-79-3 .

REFERENCES

[1] Birmbaum, A. Some latent trait models and their use in infering an examinee's ability. In F. M. Lord & M. R. Novick. (Eds.), Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

[2] Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent theory. Psychometrika, 1973, 38, 221-233.

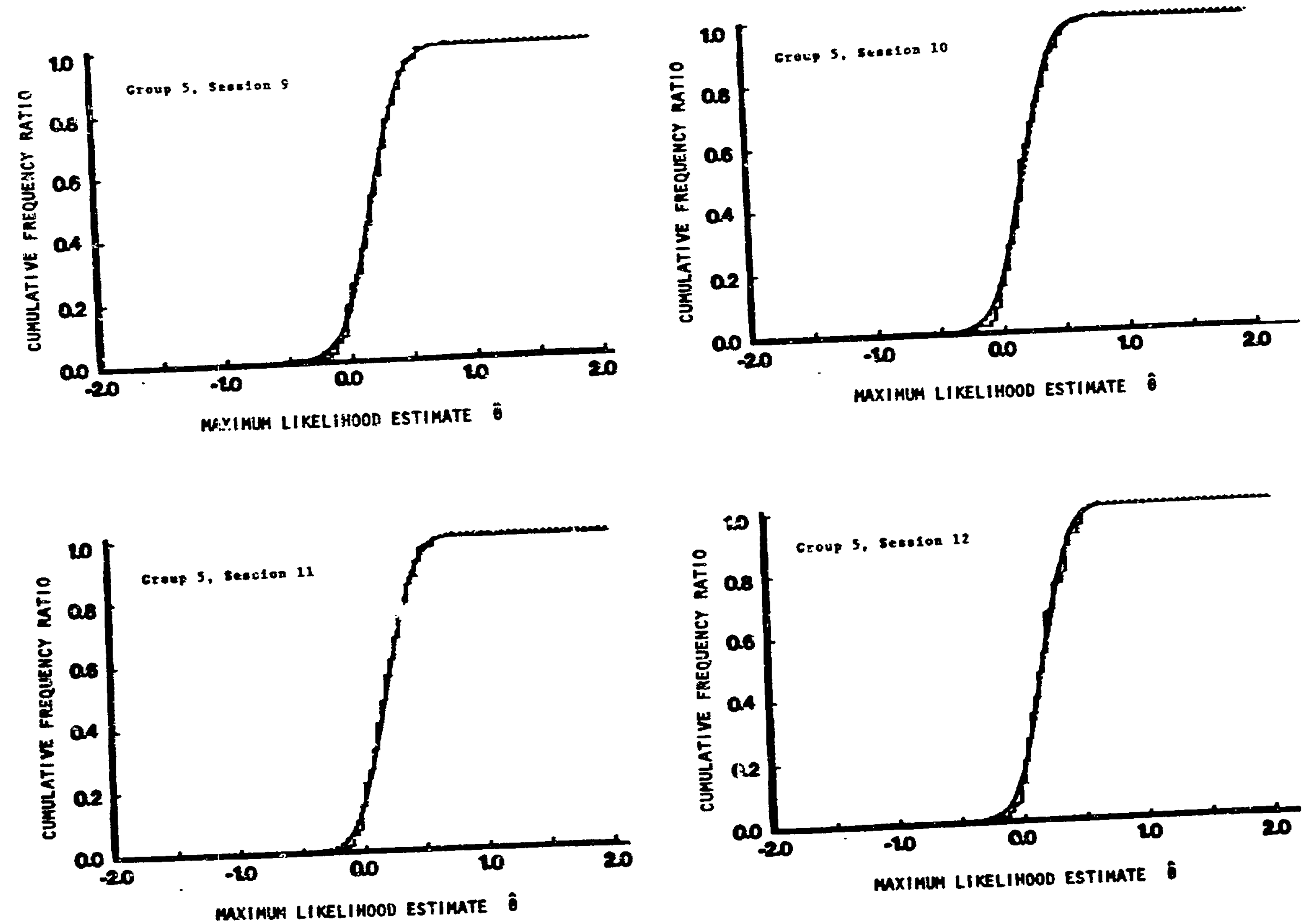


FIGURE 8-8-3

Cumulative Frequency Ratio of the Maximum Likelihood Estimate $\hat{\theta}$ of the 100 Hypothetical Examinees (Step Function) Whose Ability Levels Are Uniformly 0.2, with the Asymptotic Normal Distribution Function (Solid Curve) and the Normal Distribution Function with the Sample Mean and Standard Deviation of $\hat{\theta}$ As Its Two Parameters (Dotted Curve), After Completing Sessions 9, 10, 11, 12, 17, 18, 19 and 20, Respectively.

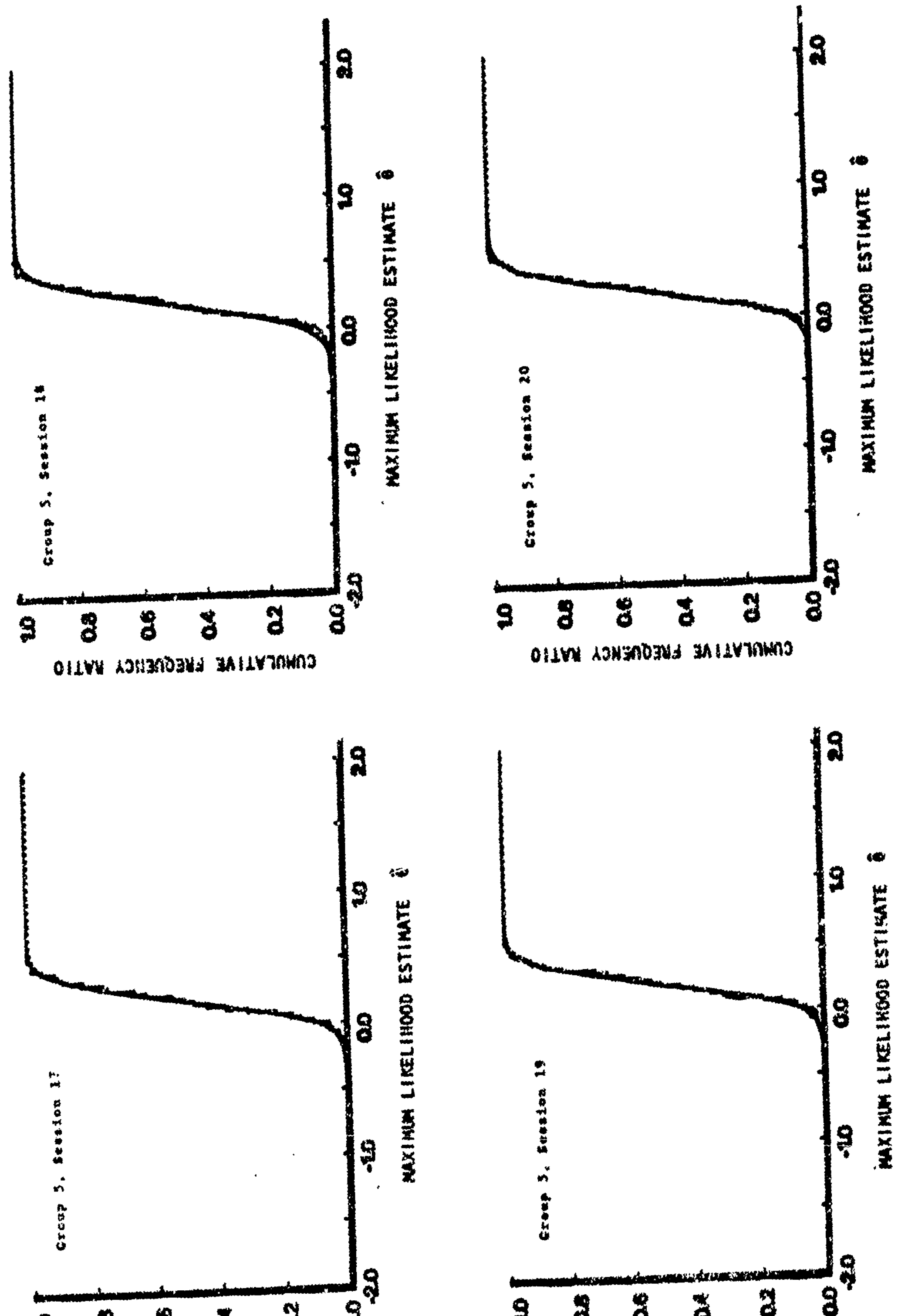


FIGURE 8-8-3 (Continued)

IX A New Family of Models for the Multiple-Choice Test Item: I

In this chapter, we shall start summarizing the rationale and findings of the part of the research, a new family of models for the multiple-choice test items, which relates to one of the main objectives of the present study. In so doing, we shall introduce the study which the author conducted in Tokyo, Japan, with the collaboration of Japanese researchers, including Dr. Sukeyori Shiba and his group of eductional psychologists and Dr. Takahiro Sato and his group of educational engineers. For simplicity, in this and next chapters, the research will be referred to as Tokyo Research.

(IX.1) Mathematical Models and Psychological Reality

Psychometricians pursue methodologies to the extent that some specific, narrowly focused topics may become their life works. This phenomenon is well exemplified in the large number of papers published in Psychometrika, which are focused upon various specific topics of factor analysis. Although it has its own merits, if we are soley satisfied with this type of research, we may overlook a more important aspect of research, i.e., psychological reality. Consequently, our work may not contribute to the progress of science to a great extent.

Mathematical models have played an important role in psychology as an science. The validation of mathematical models with psychological reality has attracted less attention from researchers, however. Needless to say, a mathematical model is nothing unless it has a sound rationale to represent our psychological reality, and, consequently, we shall be able to design and organize our research to obtain, without distortions, meaningful findings and future directions. Researchers' conscience preassumes the virtue of doubts. We cannot emphasize enough that the soundness of the rationale behind any mathematical model and its fitness to our psychological reality are by far the most important to our research. For this reason, the author has developed various methods and approaches for estimating the operating characteristics of discrete item responses without assuming any mathematical forms (cf. Chapters 3, 5 and 6). When we are not certain, we may approach the subject withour assuming any mathematical models.

(IX.2) Three Parameter Logistic Model

Three-parameter logistic model (Birnbaum, 1968) has been widely used for the multiple-choice test item among psychometricians and other researchers in mental measurement. The model is based upon the knowledge or random guessing principle, i.e., the examinee either knows the answer, or guesses randomly ard picks up an arbitrary alternative. Let $\Psi_g(\theta)$ be the item characteristic function in the logistic model, which is given by (8.12). The three-parameter logistic model is defined by the item characteristic function such that

$$P_g(\theta) = (1-c_g)\,\Psi_g(\theta) + c_g \quad , \tag{9.1}$$

where c_g is the third parameter, which is called the guessing parameter. In spite of the popularity of the model, very few researchers have tried to validate, or invalidate, the model with their own data.

It is common among experienced test constructors to include wrong, but plausible, answers among the alternatives of a multiple-choice item, which are called distractors, so as not to make its correct answer too conspicuous and destroy the quality of the question. It is noted that we need some higher mental processes other than random guessing to recognize the plausibility of a distractor, and to be attracted to it. It is contradictory, therefore, to apply the three-parameter normal ogive, or logistic, model for multiple-choice items with such distractors, although many researchers seem to like the model.

The third parameter of the three-parameter logistic model, c_g, is often called pseudo-guessing parameter, and its estimate tends to be less than unity divided by the number of the alternatives (e.g. Lord, 1968). This fact itself is the invalidation of the model, although many researchers do not admit it. It is apparent that something other than random guessing is included in our psychological reality, which makes us choose wrong answers in preference to the correct answer.

Some other model, or models, is desirable which fits our psychological reality better.

(IX.3) Tokyo Research

In the summer of 1979, the author spent a few weeks in Tokyo, Japan, with the support of the Office of Naval Research, and had conferences with researchers in Japan. The scientific monograph published in 1980 (cf. Chapter 2), with the help of Dr. Rudolph J. Marcus, Scientific Director of the ONR Tokyo Office, is based upon this research. The researchers with whom the author had conference include Dr. Takahiro Sato and Dr. Sukeyori Shiba. The author had two more opportunities to have conferences with them in the summers of 1980 and 1981 . Among others, Dr. Shiba and the author started a long term collaboration in research in 1979 , which concerns with his word comprehension tests, and mathematical models for the multiple-choice test items. It will eventually incorporate the author's methods and approaches for estimating the operating characteristics of the discrete item responses in a large scale of empirical study.

In Section IX.4 , a brief introduction to Sato's research on Index k will be made. Shiba's research and his word comprehension tests will be introduced in Sections X.1 through X.3 of the next chapter.

(IX.4) Sato's Index k

Let g $(=1,2,\ldots,n)$ be a multiple-choice test item. In this section, however, this symbol g is omitted, whenever it is clear that we deal with only one item. Let i $(=1,2,\ldots,m)$ be an alternative, or an option, of the multiple-choice item g , and p_i be the probability with which the examinee selects the alternative i . The entropy H is defined as the expectation of $-\log_2 p_i$ such that

$$(9.2) \qquad H = -\sum_{i=1}^{m} p_i \log_2 p_i \ ,$$

for the set of m alternatives of item g . It is obvious from (9.2) that the entropy H is non-negative, and, if one of the m alternatives is the sure event with unity as its probability, then $H = 0$. Sato's Index k is defined by

$$(9.3) \qquad k = 2^{H} ,$$

and is used as an index of the effectiveness of the set of m alternatives for item g in the context of information theory. Since the entropy H indicates the expected uncertainty of the set of m events, or alternatives, the set of alternatives is more informative for a greater value of k .

When the probability p_i is replaced by the frequency ratio, P_i , we can write for the estimate of the entropy such that

$$(9.4) \qquad \hat{H} = - \sum_{i=1}^{m} p_i \log_2 p_i ,$$

and for the estimate of k we have

$$(9.5) \qquad \hat{k} = 2^{\hat{H}} .$$

We notice that we can obtain the number of hypothetical, equivalent alternatives k without using the entropy, for we have

$$(9.6) \qquad k = 2^{H} = 2^{-\sum_{i=1}^{m} p_i \log_2 p_i} = \prod_{i=1}^{m} p_i^{-p_i} = [\prod_{i=1}^{m} p_i^{p_i}]^{-1} .$$

The quantity in the brackets of the last expression of (9.6) is a kind of *weighted geometric mean* of $_i$. Equation (9.6) also implies that we can use any base for $\log p_i$, instead of 2 . For convenience, hereafter we shall use e as the base of $\log p_i$, and use H^* instead of H such that

$$(9.7) \qquad H^* = - \sum_{i=1}^{m} p_i \log_e p_i \geqslant 0 ,$$

which equals zero when one of the alternatives is the sure event, and

(9.8) $$k = e^{H^*} \geqslant 1 \quad ,$$

and simply write $\log p_i$ instead of $\log_e p_i$.

To find out the value of p_i which maximizes H^* , and hence k , we define Q such that

(9.9) $$Q = - \sum_{i=1}^{m} p_i \log p_i + \lambda[\sum_{i=1}^{m} p_i - 1] \quad ,$$

where λ is Lagrange's multiplier. Thus the partial derivative of Q with respect to p_i is given by

(9.10) $$\frac{\partial Q}{\partial p_i} = -[\log p_i + (1/p_i)p_i] + \lambda = -\log p_i + (\lambda - 1) \quad .$$

Setting this derivative equal to zero, we obtain

(9.11) $$\log p_i = \lambda - 1 \quad ,$$

which is a constant regardless of the value of i . Since we have

(9.12) $$\sum_{i=1}^{m} p_i = 1 \quad ,$$

we obtain

(9.13) $$\hat{p}_i = 1/m \quad .$$

Thus it is clear that H^* , and hence k , is maximal when all the m alternatives are equally probable, and we can write

(9.14) $$\max.(H^*) = \log m$$

and

(9.15) $\max.(k) = m$.

Since in the present situation the m events are alternatives, the values of H^* and k are affected by the difficulty level of item g . Let R be the correct answer to item g , which is given as one of its alternatives, and p_R be the probability with which the examinee selects the correct answer R . Figure 9-4-1 presents the relationship between the probability p_R and the number of hypothetical, equivalent alternatives k . In this figure, the area marked by slanted lines indicates the set of k 's which are less than $\max.(k|p_R)$ and greater than $\max.[1/p_R, \min.(k|p_R)]$, and are considered to be reasonable values of k by Sato and others. In practice, Figure 9-4-1 is used by replacing the probability p_R by

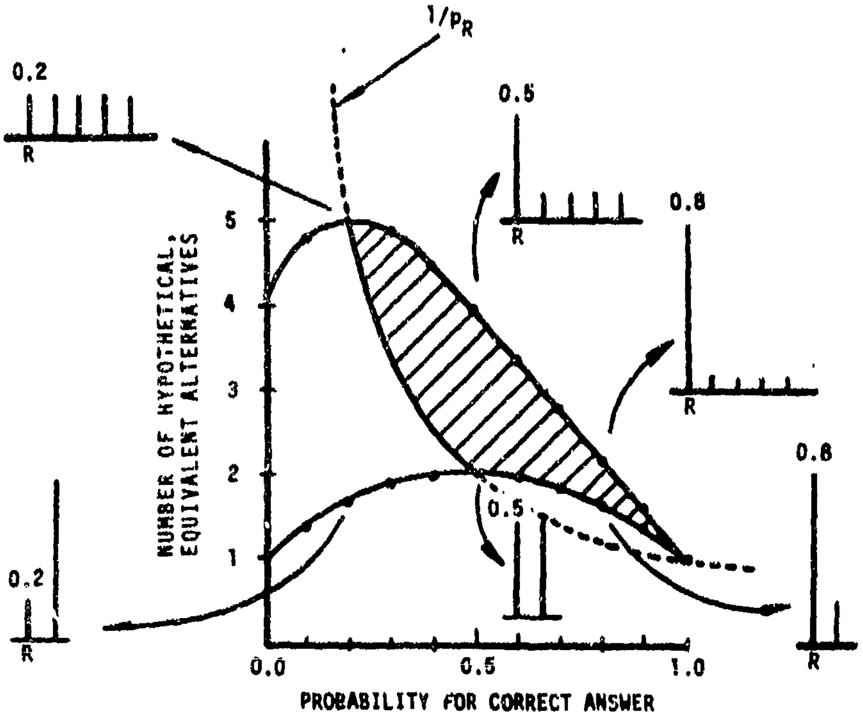


FIGURE 9-4-1

Relationship between the Probability with Which the Correct Answer R Is Selected and the Number of Hypothetical, Equivalent Alternatives, for Five-Choice Items. (Sato's Data)

the proportion correct, P_R , and the number of hypothetical, equivalent alternatives, k , by its estimate $\hat{k}$.

(IX.5) Index k* for the Validation Study of the Three-Parameter Logistic Model

Sato's Index k takes on a high value, if every examinee in the group has selected one of the m alternatives at random. This fact implies that, although the index was introduced for quite an opposite purpose and proved its usefulness, it may also be useful in detecting the examinee's random guessing behavior in quite a different situation, i.e., the multiple-choice testing. In so doing, it will be more convenient if we can modify Sato's Index k in such a way that it is unaffected by the ability distribution of a specific population of examinees, and can be considered as a pure property of the item. With this aim in mind, we shall introduce a new index, i.e., Index k^* .

Let $\bar{A}$ be the event that the examinee does not know the answer to item g , and consider the probability space which consists of such a subpopulation of examinees. The conditional probability, $p(i|\bar{A})$, with which the examinee selects the alternative i of item g in this conditional probability space is given by

$$(9.16) \qquad p(i|\bar{A}) \begin{cases} = p_i[\sum_{i\neq R} p_i + p_R^*]^{-1} & i\neq R \\ = p_R^*[\sum_{i\neq R} p_i + p_R^*]^{-1} , & i=R \end{cases}$$

where p_R^* denotes the probability with which the examinee guesses correctly for item g . The new index, k^* , is defined in terms of these conditional probabilities, in such a way that

$$(9.17) \qquad k^* = \exp[-\sum_{i=1}^{m} p(i|\bar{A})\cdot\log p(i|\bar{A})] = [\prod_{i=1}^{m} p(i|\bar{A})^{p(i|\bar{A})}]^{-1} .$$

It is obvious that $p(i|\bar{A})$ for $i\neq R$ is proportional to p_i , for every examinee in the population who has selected one of the wrong answers does not know the answer, and consequently, he is also in the subpopulation $\bar{A}$. On the other hand, examinees who have selected the correct answer R are not necessarily in the suspopulation $\bar{A}$, so we can write

$$(9.18) \qquad p_R^* \leqslant p_R \ .$$

Note that, if the examinee's behavior follows the knowledge or random guessing principle and the item characteristic function of the multiple-choice item g is of one of the three-parameter models, p_R^* equals p_i for $i \neq R$, and, as the result, all the m $p(i|\bar{A})$'s are equal and $k^* = m$.

In practice, we need to use some estimates for $p(i|\bar{A})$'s , to obtain the estimate of k^* . Since we have the frequency ratio, P_i , for the estimate of p_i for $i \neq R$, all we need to do is to find out an appropriate estimate of p_R^* . Let P_R^* denote such an estimate of p_R^* , and P_i^* be such that

$$(9.19) \qquad P_i^* \begin{cases} = P_i & i \neq R \\ = P_R^* & i = R \ . \end{cases}$$

Then we can write for the estimate of $p(i|\bar{A})$ such that

$$(9.20) \qquad \hat{p}(i|\bar{A}) = P_i^* [\sum_{i=1}^{m} P_i^*]^{-1} \ .$$

We are to take the strategy of finding P_R^* which makes k^* maximal. Define $\hat{H}^*$ such that

$$(9.21) \qquad \hat{H}^* = \log \hat{k}^* = - \sum_{i=1}^{m} \hat{p}(i|\bar{A}) \cdot \log \hat{p}(i|\bar{A})$$

$$= -[\sum_{s=1}^{m} P_s^*]^{-1} [\sum_{i=1}^{m} P_i^* \cdot \log P_i^* - (\sum_{i=1}^{m} P_i^*) \cdot \log \{\sum_{s=1}^{m} P_s^*\}] \ .$$

Then the partial derivative of $\hat{H}^*$ with respect to P_R^* can be written as

$$(9.22) \qquad \frac{\partial \hat{H}^*}{\partial P^*_R} = [\sum_{s=1}^{m} P^*_s]^{-2} [\sum_{i=1}^{m} P^*_i \cdot \log P^*_i - (\sum_{s=1}^{m} P^*_s) \cdot \log P^*_R] \quad ,$$

and, setting this equal to zero, we obtain

$$(9.23) \qquad \log P^*_R = [\sum_{s \neq R} P_s]^{-1} \sum_{s \neq R} P_i \cdot \log P_i \quad ,$$

and then

$$(9.24) \qquad P^*_R = \prod_{i \neq R} P_i^{P_i [\sum_{s \neq R} P_s]^{-1}} \quad .$$

Thus we can use (9.24) in (9.19), and, therefore, obtain $\hat{p}(i|\bar{A})$ through (9.20). The estimate of the new index, k^* , is given by

$$(9.25) \qquad \hat{k}^* = \exp[-\sum_{i=1}^{m} \hat{p}(i|\bar{A}) \cdot \log \hat{p}(i|\bar{A})] = [\prod_{i=1}^{m} \hat{p}(i|\bar{A})^{\hat{p}(i|\bar{A})}]^{-1} \quad .$$

A necessary, though not sufficient, condition for one of the three-parameter models to be valid is that $\hat{k}^*$ should be equal to m within sampling fluctuations, regardless of the population of examinees from which our sample happened to be selected. If this is not the case, we must say that the three-parameter model does not fit our item, i.e., the invalidation of the model.

(IX.6) Simulation Study on Index k^*

For the purpose of illustration, a set of simulated data was calibrated, using the Monte Carlo method. In this set of data, five hypothetical multiple-choice test items were assumed, each having five alternatives, A, B, C, D and E, with A always as the correct answer. Each item is assumed to follow the three-parameter normal ogive model, and its parameter values are shown in Table 9-6-1. A group of five hundred hypothetical examinees was

TABLE 9-6-1

Item Discrimination Parameter a_g and Item Difficulty Parameter b_g of Each of the Five Hypothetical, Binary Items Following the Three-Parameter Normal Ogive Model, with $c_g = 0.2$.

Item	a_g	b_g
1	1.00	0.00
2	1.50	0.00
3	2.00	0.00
4	2.50	0.00
5	3.50	0.00

assumed, whose ability levels are placed at one hundred equally spaced points on the ability continuum, which start with -2.475 and end with 2.475 , in such a way that subjects 1 through 5 are placed at $\theta = -2.475$, subjects 6 through 10 are at $\theta = -2.425$, and so on. For each of the five hypothetical multiple-choice items, the response of each of the five hundred hypothetical examinees was calibrated according to the specified item characteristic function with the knowledge or random guessing principle.

Table 9-6-2 presents the frequency ratio, p_i , of each of the five alternatives, for each of the five hypothetical multiple-choice items. We can see that sampling fluctuations are fairly large for item 4, and to a less degree for item 2, since the corresponding probability, p_i , is 0.6 for the alternative A and 0.1 for each of the alternatives B, C, D and E . In the same table, also presented are the values of P^*_R , which were obtained through (9.24). Using these values in (9.21), (9.24) and (9.25), the estimates of the entropy H^* and the index k^* were obtained, and are presented in Table 9-6-3. Since the maximal possible value of $\hat{H}^*$ is approximately 1.60944 ($=\log m$) and that of $\hat{k}^*$ is 5 ($=m$) , we can say that these results are sufficiently close to their respective maximal values, i.e.,

TABLE 9-6-2

Frequency Ratio of the Subject, P_i , Who Selected Each of the Five Alternatives, and the Modified Frequency Ratio P_R^* for the Correct Answer A, for Each of the Five Hypothetical Items.

Alternative Item		A	B	C	D	E
1	P_i	.608	.086	.106	.100	.100
	P_R^*	.098				
2	P_i	.618	.102	.080	.106	.094
	P_R^*	.096				
3	P_i	.600	.094	.106	.108	.092
	P_R^*	.100				
4	P_i	.606	.104	.078	.130	.082
	P_R^*	.101				
5	P_i	.598	.092	.100	.104	.106
	P_R^*	.101				

TABLE 9-6-3

Entropy, $\hat{H}^*$, and the Number of Hypothetical, Equivalent Alternatives, $\hat{k}^*$, for Each of the Five Hypothetical Items Following the Three-Parameter Normal Ogive Model.

Item	$\hat{H}^*$	$\hat{k}^*$
1	1.60714	4.98853
2	1.60501	4.97789
3	1.60744	4.99000
4	1.59224	4.91475
5	1.60829	4.99424

an exemplification of the satisfaction of one of the necessary conditions for validating the three-parameter normal ogive model and the knowledge or random guessing principle by our simulated data. The fact that these results are less satisfactory for item 4 and that the same is true, to a lesser degree, for item 2 must be due to the sampling fluctuations, which were observed in Table 9-6-2.

For the detail of this study, the reader is directed to ONR-Tokyo Scientific Monograph 3, Chapter 5.

(IX.7) Iowa Tests of Basic Skills

Concerning the validation of mathematical models, an empirical study was conducted using test data provided by Dr. William Coffman of the University of Iowa, who is also Director of the Iowa Testing Programs. For simplicity, hereafter, we shall call them Iowa Data, and this part of research Iowa Study. The data analysis of this part of the research was conducted by the persistent effort of one of the author's assistants, Robert Trestman.

The battery of tests used here is the Iowa Tests of Basic Skills, Form 6, Levels 9-14. These tests have been designed, constructed, and revised at the College of Education of the University of Iowa since 1935, with the general school population in mind, and for students of ages nine through fourteen, or grades three through nine. All technical information in this paper has been taken from either Form 6 itself (Hieronymous and Lindquist, 1971), or its Teacher's Manual (Iowa Basic Skills Testing Program, 1971).

There are eleven tests in the battery, each of which focuses on a different basic skill. For convenience, hereafter, we shall call these separate test· subtests, in order to avoid the confusion which might occur when we refer to both the total test battery and each test in the battery. Following the Teacher's Manual, the descriptions and abbreviations of these eleven subtests, together with their administration schedule and working times, are tabulated and presented in Table 9-7-1. All the test items are power test items with multiple-choice format, with five alternative answers for the items in

TABLE 9-7-1

Administration Sessions, Time Limits and Subtests of Iowa Tests of Basic Skills.

Administration Session	Working Time (Minutes)	Subtest
First Session 85 Minutes	17 55	V: Vocabulary R: Reading Comprehension
Second Session 80 Minutes	12 15 20 20	L-1: Spelling L-2: Capitalization L-3: Punctuation L-4: Usage
Third Session 85 Minutes	30 20 30	W-1: Map Reading W-2: Reading Graphs and Tables W-3: Knowledge and Use of Reference Materials
Fourth Session 65 Minutes	30 30	M-1: Mathematics Concepts M-2: Mathematics Problem Solving

Subtest L1, and with four alternatives for those in the other ten subtests. These eleven subtests are designed to cover all major areas of academic interest for the grades three through nine.

The numbers of test items contained by the eleven separate subtests are 114, 178, 114, 102, 102, 86, 89, 74, 141, 136 and 96, respectively, following the order of subtests given in Table 9-7-1. For each of the five levels, 9 through 14, only a subset of each subtest is administered. The standardized administration schedule and the working time for each subtest are presented in Table 9-7-1. For the entire test battery, the time required for the administration of each level of test is four hours and thirty-nine minutes. It is recommended that the test be administered on four consecutive days.

In our data, only the tests of Levels 11, 12 and 13 were used. The numbers of test items contained in these three levels of test are 461, 487 and 500, respectively. A graphical representation is made in Figure 9-7-1, to show how these three subsets of test items in each subtest overlap among the three levels.

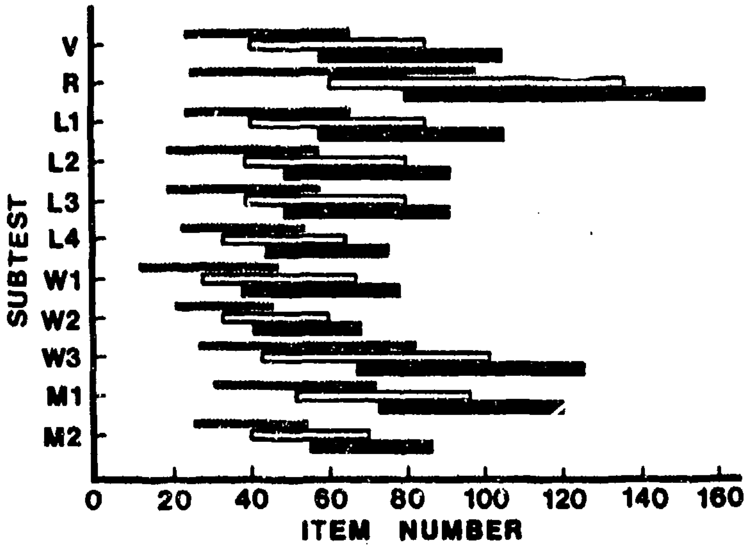


FIGURE 9-7-1

Test Items of Each of the Eleven Subtests of Iowa Tests of Basic Skills Administered to Each of Levels 11, 12 and 13 , Which Are Represented by Shaded, Hollow, and Solid Bars, Respectively.

We notice in Figure 9-7-1 that all the test items given to the students of Level 12 are also given to those of Level 11 or Level 13, or both. There are exactly one hundred test items which are given to all the three levels of examinees. There are 264 which are given to Levels 11 and 12, and 323 to Levels 12 and 13, respectively. We also have 197 items which are taken by the examinees of Level 11 only, and 177 by those of Level 13 only. Thus the total number of test items is 1,061.

(IX.8) Original and Revised Iowa Data

Data were collected in three different school systems in the State of Iowa, in the years 1971 through 1977. In their original form, the total number of examinees, including both boys and girls, is 7,581. Out of these people, 28 students took Level 9 Test and 114 took Level 10 Test. Since these are relatively small numbers, we decided to exclude them from our original group of examinees. The other 7,439 examinees are classified into three subgroups, i.e., 2,460 students who took Level 11 Test, 2,452 who took Level 12 Test, and 2,527 who took Level 13 Test. Hereafter, we shall call observations concerning these 7,439 examinees the original data.

It was found out that there are a relatively small number of examinees who did not respond to a substantially large number of test items. While as many as 7,010 examinees out of the total 7,439 examinees left only 49 or less test items unanswered, there also are 162 examinees who did not respond to as many as 100, or more, test items. Our raw data show there are some examinees included who skipped an entire subtest, or more than one entire subtest. A close examination of the original data indicates that, if we exclude all the examinees who left, at least, one half of a subtest unanswered from our total group of examinees, tnen the number of examinees who left 200 or more test items unanswered will become zero, and only 28 examinees, who omitted more than 100, but less than 200, test items, will be included. For this reason, we have decided to exclude the 193 examinees who left one half of a subtest, or more, unanswered from our original group of examinees for the detailed analysis. Hereafter, we shall call observations concerning the remaining 7,246 examinees the revised data, to distinguish them from the original data.

Table 9-8-1 presents the item identifications of the fifty-five test items, i.e., 34 for Level 11, 15 for Level 12, and 6 for Level 13, to which less than 90 percent of examinees responded in the original data, their percentages in the original and revised data, respectively. We can see in this table that for most of these fifty-five test items the two percentages show a visible improvement caused by the exclusion of the 193 examinees. There is a substantial improvement in the percentage of examinees who answered in one way or another, for all the three levels, which was provided by the exclusion of the 193 examinees. Among others, we notice that the frequency of test items which were answered by 99 percent, or more, of examinees increased from 231 to 320 for Level 11, from 319 to 350 for Level 12, and from 286 to 377 for Level 13.

TABLE 9-8-1

Fifty-Five Test Items of Iowa Tests of Basic Skills to Which Less Than Ninety Percent of Examinees Responded in One of the Three Levels in the Original Data, the Percentages of Responses in the Original Data, and Those in the Revised Data.

Item	Level 11		Level 12		Level 13	
	Original	Revised	Original	Revised	Original	Revised
V-66	89.1	91.4				
R-95	89.6	91.9				
R-96	89.2	91.5				
R-97	88.7	91.0				
R-98	88.3	90.7				
L1-62	89.4	91.8				
L1-63	88.3	90.7				
L1-64	87.5	90.0				
L1-65	86.3	88.7				
L1-66	84.8	87.3				
L1-80			89.7	90.7		
L1-81			88.9	89.9		
L1-82			87.9	88.9		
L1-83			87.0	88.0		
L1-84			86.2	87.2		
L1-85			85.4	86.4		
L1-105					89.7	90.6
W1-41	88.9	91.4				
W1-42	85.6	88.2				
W1-43	83.3	86.0				
W1-44	81.7	84.3				
W1-45	79.2	81.6				
W1-46	76.7	79.0				
W1-47	74.7	76.9				
W1-60			89.2	90.4		
W1-61			87.2	88.3		
W1-62			85.2	86.3		
W1-63			82.7	83.8		
W1-64			80.8	81.9		
W1-65			78.4	79.4		
W1-66			75.4	76.3		
W1-67			74.2	75.2		
W1-74					88.9	90.0
W1-75					87.3	88.4
W1-76					86.2	87.2
W1-77					85.0	86.1
W1-78					83.9	84.9
W3-69	90.0	92.3				
W3-70	89.3	91.5				
W3-71	88.8	91.0				
W3-72	87.9	90.2				
W3-73	87.1	89.4				
W3-74	86.8	89.0				
W3-75	86.0	88.3				
W3-76	84.9	87.2				
W3-77	84.0	86.3				
W3-78	83.5	85.8				
W3-79	82.8	85.0				
W3-80	82.3	84.5				
W3-81	81.6	83.8				
W3-82	81.1	83.3				
M2-52	87.9	90.0				
M2-53	85.9	87.9				
M2-54	83.7	85.7				
M2-69			89.4	90.3		

Table 9-8-2 presents the frequency distribution of test items for each of the eleven subtests with respect to the percentage of examinees who answered correctly, for each of Levels 11, 12 and 13, for the revised data. It should be noted that, even in the revised data, these percentages correct are not independent from the positions of the test items in each subtest. There is a distinct tendency that larger numbers of examinees did not respond to items which were presented later in each subtest. It is obvious, therefore, that, for these later items, the percentage for the correct answer is less than it should be in the ideally set free-response situation.

(IX.9) Informative Distractor Model

By Informative Distractor Model, we mean the family of models in which we assume the existence of specific information obtainable from separate alternative answers, including the correct answer, of each multiple-choice test item.

TABLE 9-8-2

Frequency Distribution of Items for Each of the Eleven Subtests with Respect to the Percentage of Examinees Answering Correctly. Each Interval of Percentage Is Greater than or Equal to the Lower End and Less than the Upper End.

Iowa Revised Data, Level 11

	Percentage	Subtest											Total
		V	R	L1	L2	L3	L4	W1	W2	W3	M1	M2	
1	0.0 - 5.0												0
2	5.0 - 10.0												0
3	10.0 - 15.0												0
4	15.0 - 20.0		1	2	1								4
5	20.0 - 25.0		1	1	1		1	1			1	1	7
6	25.0 - 30.0	1	2				1	1				4	9
7	30.0 - 35.0	1	4	2			1	2		1	3	2	16
8	35.0 - 40.0	3	2	2	1	2	1	2		1	4	3	21
9	40.0 - 45.0	4	5	6	3	4	5	1	2	3	3	2	38
10	45.0 - 50.0	4	9	7	6	5	6	1	3	4	3		48
11	50.0 - 55.0	4	8	3	3		4	4	7	10	8	3	54
12	55.0 - 60.0	10	5	5	2	5	4	6	3	15	3	2	60
13	60.0 - 65.0	5	4	1	9	5	5	4	2	6	2	3	46
14	65.0 - 70.0	4	6	3	5	9	3	6	3	9	5	4	57
15	70.0 - 75.0	4	9	5	3	2	1	2	1	6	3	3	39
16	75.0 - 80.0	2	5	6	4	6		1	1	1	4	2	32
17	80.0 - 85.0	1	7		1	1		2	2		1		15
18	85.0 - 90.0		4		1	1		1	1		1		9
19	90.0 - 95.0		2					2	1		1		6
20	95.0 -100.0												0
	Total	43	74	43	40	40	32	36	26	56	42	29	461

TABLE 9-8-2 (Continued)

Iowa Revised Data, Level 12

	Percentage	Subtest											Total
		V	R	L1	L2	L3	L4	W1	W2	W3	M1	M2	
1	0.0 - 5.0												0
2	5.0 - 10.0												0
3	10.0 - 15.0												0
4	15.0 - 20.0					1							1
5	20.0 - 25.0	1	1	1	1	1					2	1	8
6	25.0 - 30.0		1	5	1		2	1		1	2	2	15
7	30.0 - 35.0	1	3	5			2	3	1	3	1	1	20
8	35.0 - 40.0	6	6	5	2	3	2	2	1	4	2	4	37
9	40.0 - 45.0	1	2	2	3	1	1	4	4	4	4	3	29
10	45.0 - 50.0	3	10	4	1	6	6	7	1	5	4	5	52
11	50.0 - 55.0	4	12	4	2	8	5	4	1	2	7	1	50
12	55.0 - 60.0	7	9	6	5	6	5	4	4	6	4	5	61
13	60.0 - 65.0	8	12	5	6	4	5	2	6	1	4	1	54
14	65.0 - 70.0	4	5	4	5	5	3	1	2	10	4		43
15	70.0 - 75.0	5	6	1	5	2	1	7	4	8	3	5	47
16	75.0 - 80.0	4	4	2	6	4		4	2	5	4	3	38
17	80.0 - 85.0	2	2	2	4				1	8	2		21
18	85.0 - 90.0		2			1		1	1	2	2		9
19	90.0 - 95.0		1		1								2
20	95.0 -100.0												0
	Total	46	76	46	42	42	32	40	28	59	45	31	487

Iowa Revised Data, Level 13

	Percentage	Subtest											Total
		V	R	L1	L2	L3	L4	W1	W2	W3	M1	M2	
1	0.0 - 5.0												0
2	5.0 - 10.0												0
3	10.0 - 15.0		1										1
4	15.0 - 20.0					1	1		1				3
5	20.0 - 25.0			3		2		1	1	1		4	12
6	25.0 - 30.0	1	2	3	1	2	1	1		1		5	17
7	30.0 - 35.0	4	3	4	2	1	5	1	2	1	6	2	31
8	35.0 - 40.0	3	4	6	1	3	3	5	3	6	3		37
.	40.0 - 45.0	4	11	11	2	2	3	3	2	8	8	3	57
10	45.0 - 50.0	2	4	2	5	4	1	2	2	3	5	5	35
11	50.0 - 55.0	5	6	5	4	7	4	6	3	10	3	4	57
12	55.0 - 60.0	6	9	2	6	7	5	7		6	5	4	57
13	60.0 - 65.0	6	15	4	3	6	4	3	2	4	5	1	53
14	65.0 - 70.0	4	8	4	3	3	2	3	4	3	4		38
15	70.0 - 75.0	7	7	2	7	3	3	3	2	5	6	1	46
16	75.0 - 80.0	2	5	2	4			2	2	6	2		25
17	80.0 - 85.0	2	2		4	2		2	3	2		2	19
18	85.0 - 90.0	2	1		1			1	1	3	1	1	11
19	90.0 - 95.0							1					1
20	95.0 -100.0												0
	Total	48	78	48	43	43	32	41	28	59	48	32	500

For the type of tests Shiba's word comprehension tests belong to, which will be introduced in Section X.1 of Chapter 10, some specific model, or models, of the Informative Distractor Model is called for. Models A, B and C proposed by the author, which will be described in Section X.7 , belong to this family of models. If we succeed in developing appropriate multiple-choice test items which follow this type of models, then they will no longer be blurred images of the corresponding free-response test items, but will provide us with additional information from the distractors which free-response test items will never have.

(IX.10) Equivalent Distractor Model

In contrast to the Informative Distractor Model, Equivalent Distractor Model means the family of models in which no specific information is expected from separate incorrect answers, which are given as alternatives in the multiple-choice test item. Thus all the alternatives, except for the correct answer, of a given multiple-choice item are equivalent, since the information given by a specific alternative,.or distractor, is not different from the one given by each remaining wrong answer. The three-parameter logistic, or normal ogive, model belongs to this family of models. In this model, all the information provided by a given wrong answer is pure noise resulting from random guessing, and, therefore, the alternative is equivalent with any remaining wrong answer. Note, however, that this type of model, which is based upon the knowledge or random guessing principle, is not the only one included by the Equivalent Distractor Model. Suppose that the operating characteristic of each wrong answer of a given multimple-choice item includes some information about the examinee's ability, but all the operating characteristics, or plausibility curves, of the distractors are identical. In such a case, we can say that the test item should belong to the Informative Distractor Model in the sense that these distractors provide us with some information concerning the examinee's ability. On the other hand, we can also say that the item should belong to the Equivalent Distractor Model, since each distractor does not have any specific information which distinguishes it from the other distractors. For convenience, in the present paper, we shall take the second standpoint, defining the Informative Distractor Model in the narrower sense.

(IX.11) Index k* for the Invalidation of the Equivalent Distractor Model

It is obvious that Index k* , which was introduced in Section IX.5 , can be used for the invalidation of the Equivalent Distractor Model, and even as a weak support for the Informative Distractor Model. If Index k* turns out to be far less than m , then we must reject the hypothesis that our model should belong to the Equivalent Distractor Model. If it assumes a value close to m , then we shall say that Equivalent Distractor Model may be adequate. In both cases, however, Informative Distractor Model stays among the possibilities.

It is noted that the traditional chi-square test with (m-2) degrees of freedom for the goodness of fit for the frequencies of the (m-1) wrong answers with the uniform distribution as the theoretical distribution may serve our purpose just as well, without using Index k* . In our pilot study, we applied it for the original data of 7,439 examinees. The result turned out to be such that only 23, 22 and 21 test items indicate the acceptance of the respective uniform distributions, or the acceptance of Equivalent Distractor Model, for Levels 11, 12 and 13, respectively, even if we take as low a level of significance as 0.0005 . This comes from the fact that our sample sizes are so large that the chi-square test is very sensitive to small diversions from the hypothesized uniform distributions. We must question, however, if such small diversions mean anything for our purpose. If, for instance, the hypothesized uniform distribution provides us with the probability 0.15 for each of the three wrong answers and the true distribution gives 0.16 , 0.14 and 0.15 , respectively, then the detection of these small deviations, i.e., 0.01 , at most, will not make a strong basis for the rejection of the Equivalent Distractor Model.

In contrast to the chi-square test, the estimated Index k* is insensitive to the sample size, because the sampling fluctuation participates in the resulting estimate only thro'gh the computation of the proportions, P_i (cf. Section IX.5). Thus, if we wish to more or less ignore the sampling fluctuations of the proportions, then Index k* may be adopted, and these values can be comparable across different sample sizes.

(IX.12) Results Obtained by Using Index k* on Iowa Data

Table 9-12-1 presents the frequency distribution of the items of each of the ten subtests, excluding Subtest L1, which consists of five-alternative test items, with respect to the resultant values of the estimated Index k* , for each of Levels 11, 12 and 13. The corresponding result for Subtest L1 is presented, separately, as Table 9-12-2, for all the three levels. We can see in Table 9-12-1 that the configurations of these frequencies are similar across the three levels, with the range of the estimated Index k* , 2.25 through 4.00 , for each level. This is also the case with Subtest L1, with the range of the estimated Index k* , 2.25 through 4.50 , for most items, as is shown in Table 9-12-2. We notice in Table 9-12-1 that, for each level, the mode of the total frequency distribution is the highest category, 3.75 through 4.00 . If we examine the frequency distributions of separate subtests, however, we will notice that there are some variations among their configurations. Above all, it is noted that Subtests L2, L3 and L4 have different modes from the highest category, i.e., mostly either the category, 3.00 through 3.25 , or the category, 3.25 through 3.50 . This tendency is also shared by Subtest L1, which has five-alternative multiple-choice test items, as is shown in Table 9-12-2.

Eight examples of the frequency distribution of the examinees with respect to their choices of an answer out of the four alternatives are presented as Figure 9-12-1. These test items are selected from the subset of 76 test items for Level 13, whose Index k* 's are 3.9 or greater. For Levels 11 and 12, there are 78 and 73 such test items, respectively. In each histogram, also drawn by a dotted line is the estimated proportion, P^*_R , multiplied by the number of examinees who answered the item in one way or another, or the total number of examinees subtracted by the number of those who did not

TABLE 9-12-1

Frequency Distribution of Four-Alternative Items with Respect to Index k* for Each of the Ten Subtests of Iowa Tests of Basic Skills. The Range of Index k* Is Greater Than or Equal to the Lower End and Less Than the Upper End of Each Interval, for Each of Levels 11, 12 and 13 .

Level 11

	Range of Index k*	Subtest										Total
		V	R	L2	L3	L4	W1	W2	W3	M1	M2	
1	1.00 - 1.25											0
2	1.25 - 1.50											0
3	1.50 - 1.75											0
4	1.75 - 2.00											0
5	2.00 - 2.25											0
6	2.25 - 2.50		1	2				1				4
7	2.50 - 2.75	1	2	7	1		1			1		13
8	2.75 - 3.00	3	2	6	8	3		2	1			25
9	3.00 - 3.25	6	6	10	12	8	1	2	5	3	1	54
10	3.25 - 3.50	3	13	8	7	12	4	1	9	4	1	62
11	3.50 - 3.75	11	13	6	8	6	7	4	12	7	12	86
12	3.75 - 4.00	19	37	1	4	3	23	16	29	27	15	174
	Total	43	74	40	40	32	36	26	56	42	29	418

Level 12

	Range of Index k*	Subtest										Total
		V	R	L2	L3	L4	W1	W2	W3	M1	M2	
1	1.00 - 1.25											0
2	1.25 - 1.50											0
3	1.50 - 1.75											0
4	1.74 - 2.00											0
5	2.00 - 2.25											0
6	2.25 - 2.50		1	4								5
7	2.50 - 2.75	2	1	8	1	1				1		14
8	2.75 - 3.00	2	4	6	8	3			5	2		30
9	3.00 - 3.25	4	10	7	8	8	2		8	6	1	54
10	3.25 - 3.50	6	9	8	10	11	4	3	11	5	3	70
11	3.50 - 3.75	10	18	5	11	5	6	8	16	8	10	97
12	3.75 - 4.00	22	33	4	4	4	28	17	19	23	17	171
	Total	46	76	42	42	32	40	28	59	45	31	441

Level 13

	Range of Index k*	Subtest										Total
		V	R	L2	L3	L4	W1	W2	W3	M1	M2	
1	1.00 - 1.25											0
2	1.25 - 1.50											0
3	1.50 - 1.75											0
4	1.75 - 2.00											0
5	2.00 - 2.25											0
6	2.25 - 2.50	2		3								5
7	2.50 - 2.75	3		7	2	1			1			14
8	2.75 - 3.00	2	5	7	4	2			2	2		24
9	3.00 - 3.25	1	5	10	11	7			7	4	1	46
10	3.25 - 3.50	11	7	8	10	10	5		10	9	5	75
11	3.50 - 3.75	10	24	5	11	6	7	10	21	12	7	113
12	3.75 - 4.00	19	37	3	5	6	29	18	18	21	19	175
	Total	48	78	43	43	32	41	28	59	48	32	452

TABLE 9-12-2

Frequency Distribution of Five-Alternative Items of Subtest L1 of Iowa Tests of Basic Skills, with Respect to Index k* , for Levels 11, 12, and 13 , Respectively.

	Range of Index k*	Level			Total
		11	12	13	
1	1.00 - 1.25				0
2	1.25 - 1.50				0
3	1.50 - 1.75				0
4	1.75 - 2.00				0
5	2.00 - 2.25				0
6	2.25 - 2.50	4	4	1	9
7	2.50 - 2.75	5	1	2	8
8	2.75 - 3.00	4	4	4	12
9	3.00 - 3.25	2	2	8	12
10	3.25 - 3.50	5	11	6	22
11	3.50 - 3.75	9	7	5	21
12	3.75 - 4.00	4	5	11	20
13	4.00 - 4.25	5	6	4	15
14	4.25 - 4.50	4	5	4	13
15	4.50 - 4.75		1		1
16	4.75 - 5.00	1		3	4
	Total	43	46	48	137

answer the item at all. We can see in this figure that most of these histograms are close to rectangles, if we replace the frequency for the correct answer by the height indicated by the dotted line in each histogram.

In the total set of 227 test items, whose values of the estimated Index k* are greater than 3.9 , we find only four test items from Subtests L2, L3 and L4, i.e., L2-58 (k*=3.95473) and L3-49 (k*=3.95320) of Level 11, and L3-49 (k*=3.95658) and L2-58 (k*=3.95318) of Level 12, which are actually two items shared by both Levels 11 and 12. A close examination of the contents of the test items of these four subtests, including Subtest L1, and their results of analysis reveals the following facts.

(1) All the questions in these four language skill subtests are in the form of having the examinee find mistakes in spelling,

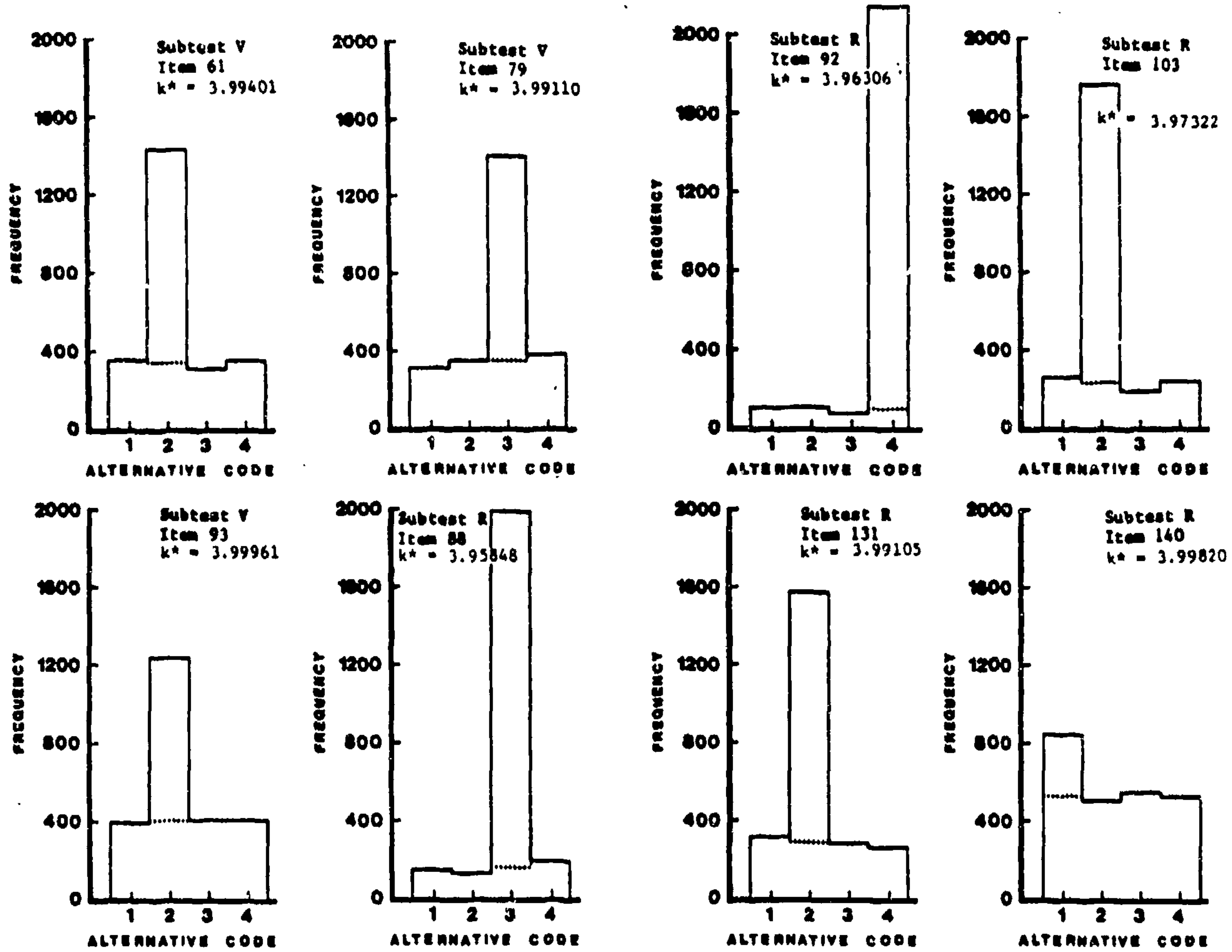


FIGURE 9-12-1

Frequency Distribution of Examinees of Level 13 with Respect to Their Responses to Each of Eight Test Items of Iowa Tests of Basic Skills Sampled from Those Whose Values of Index k* are 3.9 or Greater, with the Estimated Proportion of the Examinees Guessing Correctly (Dotted Line).

capitalization, punctuation and usage, respectively.

(2) Unlike the test items in the other seven subtests, these items have "No mistakes" as the last alternative, and for most items this alternative has a high frequency, even when it is a wrong answer.

From these facts and the above results, it is obvious that Equivalent Distractor Model is not suitable for the items of the four subtests of language skills, including Subtest L1, which consists of five-alternative test items. For these items, Informa·:ive Distractor Model may be more appropriate.

Figure 9-12-2 presents similar histograms to those in Figure 9-12-1 for the frequency distributions of eight four-alternative test items, which were selected from the subset of 9 test items whose Index k* 's are less than, or equal to, 2.6 , for Level 13. The corresponding numbers of test items are 7 for Level 11, and 11 for Level 12, respectively. We can see in this figur: that these histograms, whose frequencies for the correct answers are replaced by the corresponding dotted lines, are far from rectangles. There is no reason to accept Equivalent Distractor Model for these test items.

(IX.13) Comparison of the Results on Common Test Items for Three Levels of Examinees in Iowa Study

There are certain test items which are included in all the three levels. Their numbers are nine for Subtest V, nineteen for Subtest R, nine for Subtest L1, ten for Subtest L2, ten for Subtest L3, eleven for Subtest L4, ten for Subtest W1, six for Subtest W2 and sixteen for Subtest W3, which make the total number of test items shared by all the three levels one hundred. There is no item which is included in all three levels for Subtests M1 and M2.

It is evident that, for the behavior of the test item to follow Equivalent Distractor Model, not only the value of estimated Index k* should be close to m for one level of examinees but also for all three levels. It will be worthwhile, therefore, to compare the results across the three levels for these one hundred test items which are included in all the three levels of test. We find that only 7 out of the 91 four-alternative test items, i.e., V-61, R-88, W1-45, W1-46, W2-44, W2-45 and W3-70, have three estimates of Index k* all of which are greater than, or equal to, 3.9 . If we shift this

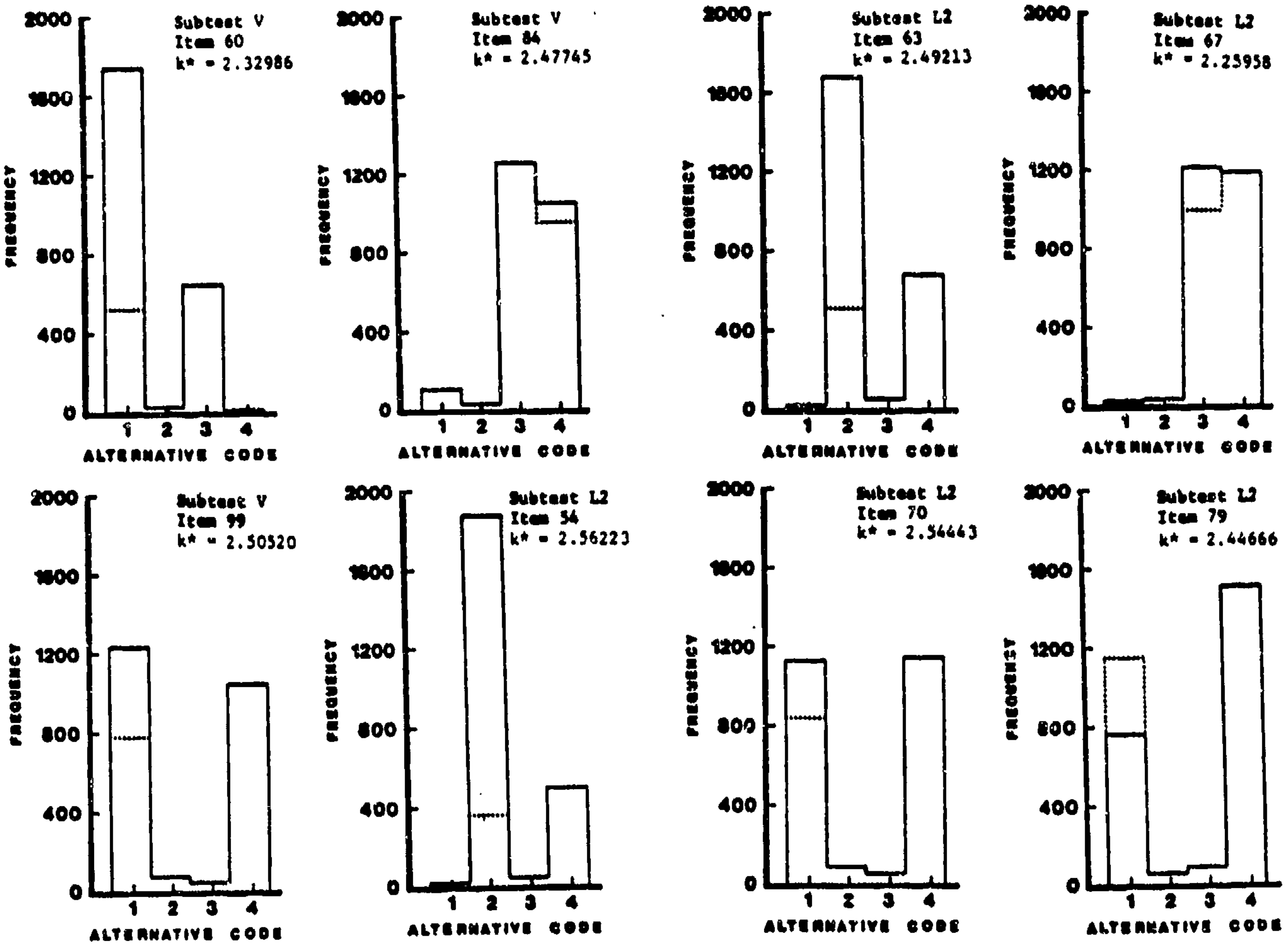


FIGURE 9-12-2

Frequency Distribution of Examinees of Level 13 with Respect to Their Responses to Each of Eight Test Items of Iowa Tests of Basic Skills Sampled from Those Whose Values of Index k* are 2.6 or less, with the Estimated Proportion of the Examinees Guessing Correctly (Dotted Line).

critical value from 3.9 to 3.8 , these seven four-alternative test items are joined by eleven more items, i.e., V-63, V-66, R-80, R-90, R-92, L2-58, L3-49, W1-40, W1-43, W1-47 and W2-41 . There are no five-alternative test items of Subtest L1 which are comparable to these eighteen four-alternative test items.

Figure 9-13-1 presents four examples of the sets of the three histograms for Levels 11, 12 and 13, which are similar to those in Figures 9-12-1 and 9-12-2, and sampled from the total nineteen shared test items of Subtest R.

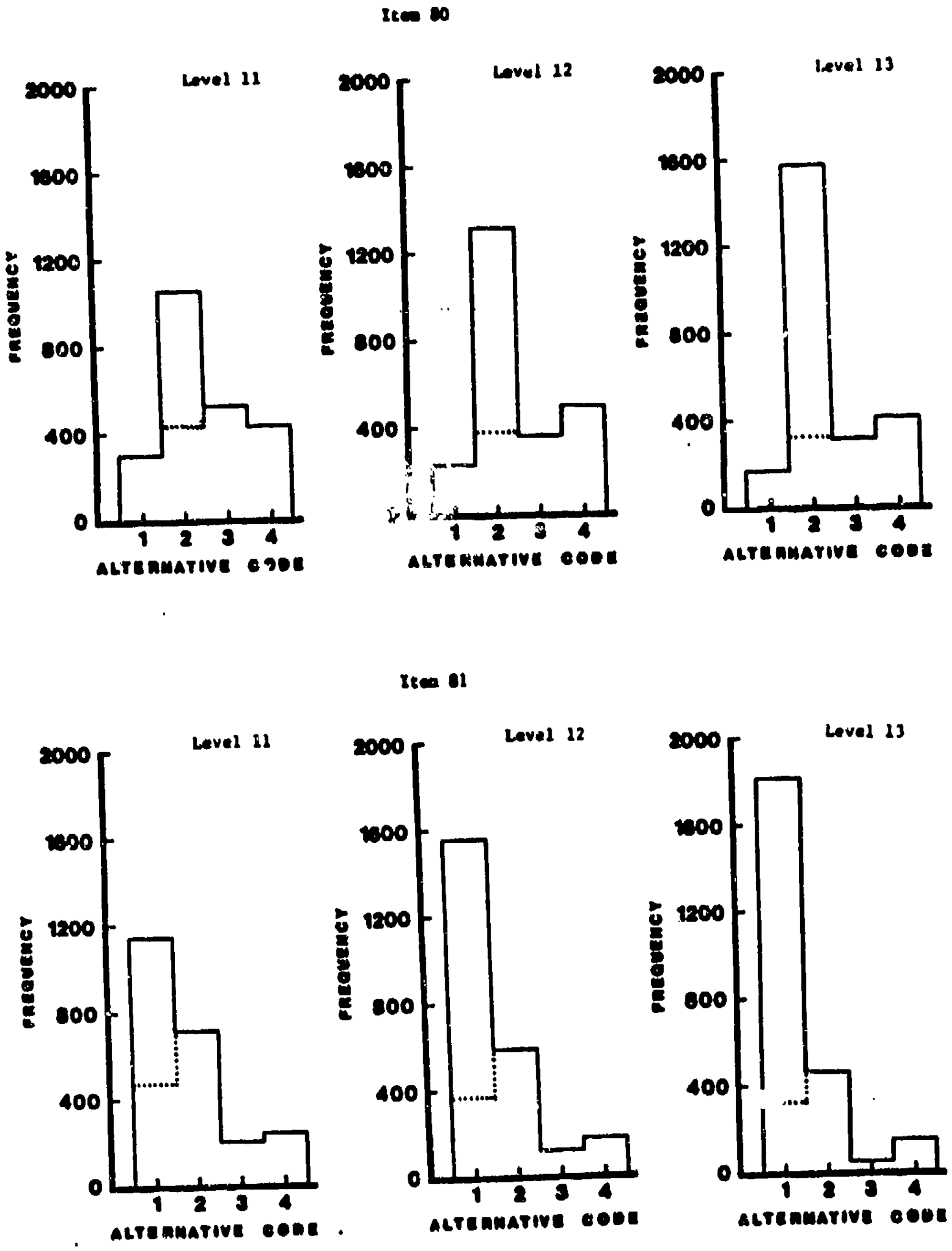


FIGURE 9-13-1

Comparison of the Three Frequency Distributions of Examinees with Respect to Their Choices of Alternatives for Each of the Twenty Items of Subtest R of Iowa Tests of Basic Skills, Which Were Administered to All Three Levels of Students, with the Estimated Proportion of Examinees Guessing Correctly (Dotted Line)

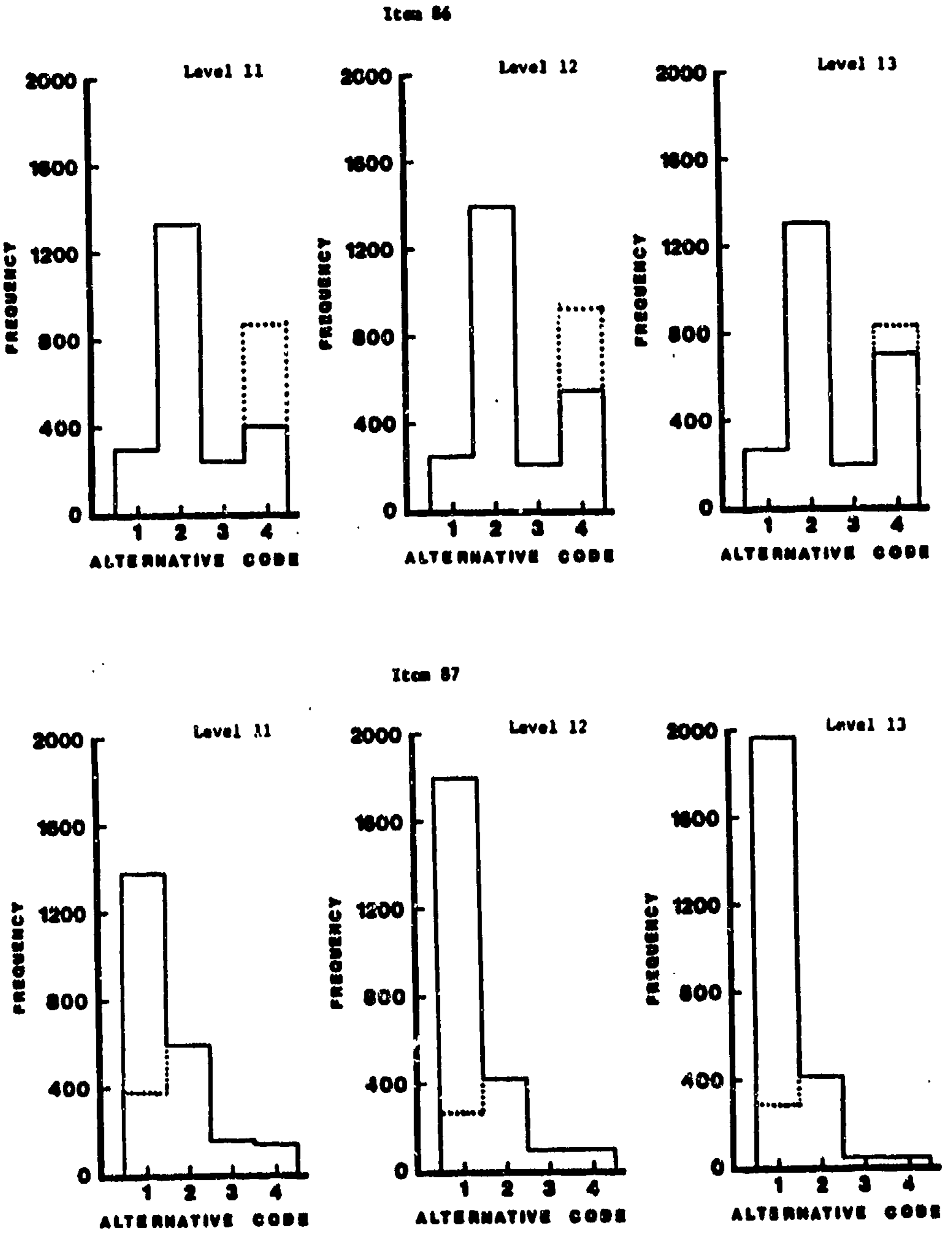


FIGURE 9-13-1 (Continued)

It is interesting to note that some items show evidence of differential information provided by separate wrong answers. For example, alternative 4 of R-80 seems to attract students of intermediate reading ability, while alternative 1 of the same item appears to attract students of lower levels of ability. It is clear that many items have one or more effective distractors, and, among others, alternative 2 of R-86 proved to be powerful. Most histograms have some regularities in the way the frequencies change across the three levels, which suggest that the examinees selected their answers intentionally rather than by random guessing.

For the detail of the Iowa Study, the reader is directed to the research report, RR-80-1.

(IX.14) Remarks on the Usage of Index k*

It should be noted that high values of Index k* can happen in situations where Informative Distractor Model is perfectly legitimate. When this happens, our information is differentiated for the separate distractors, and yet the number of examinees who selected each distractor as their answers is close to that of each other. This is an ideal situation for our purpose of mental measurement, because, not only each distractor is informative, but also all of these distractors are well used, with the examinees' answers distributing evenly over the distractors. We recall Sato's Index k, which was introduced in Section IX.4, is for this purpose, and works well in the small classroom situation where teachers supervise their students well and there is little chance for the students to make random guessing.

The above fact makes us realize that we must be careful before we make conclusions from the estimated values of Index k*. Observation of the values of Index k* across several subpopulations of examinees of different ability levels, like the one for Levels 11, 12 and 13 of Iowa Data which was introduced in the preceding section, is one of the ways of finding out the cause for high values of Index k*. If it is due to the equivalence of the distractors, then we will have similar values of Index k* across the subpopulations; if differential information exists for the separate distractors, then the values of Index k* will differ for the separate subpopulations, provided that their ability differences are substantial. Another way is to compare the sample means of the ability estimate among the subgroups of examinees who selected separate distractors for their answers. If differential information exists, then these sample means of the ability estimate will also differentiate, while they will stay close to one another if the distractors are equivalent. This was done in Shiba's study, which will be introduced in Section X.3 of the next chapter.

With these considerations in mind, Index k* can be used effectively.

REFERENCES

[1] Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M.R. Novick; Statistical theories of mental test scores. Addison-Wesley, 1968, Chapters 17-20.

[2] Hieronymous, A. N. and E. F. Lindquist. Iowa test of basic skills (Levels ed.), Form 6. Boston: Houghton-Mifflin Company, 1971.

[3] Iowa Basic Skills Testing Program. Iowa test of basic skills teacher's manual. Iowa City, Iowa: University of Iowa, 1971.

[4] Lord, F. M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

X A New Family of Models for the Multiple-Choice Test Item: II

In the preceding chapter, Index k* was introduced, and through the work on Iowa Data we have seen that the type of models, which are based upon the principle of knowledge or random guessing, do not work for many multiple-choice test items. In this chapter, Shiba's research on the word comprehension, and then the new family of models for the multiple-choice test items, will be introduced.

(X.1) Shiba's Word Comprehension Tests

The battery of tests used for the construction of the word comprehension scale consists of eleven tests, A1, A2, A3, A4, A5, A6, J1, J2, S1, S2 and U . Each test contains thirty to fifty-eight multiple-choice items, each having a set of five alternatives. These tests differ in difficulty, and each of them is designed for a different group of ages, ranging from six years of age to the ages of college students. There are subtests of items included in two tests, which are adjacent to each other in difficulty. For example, items 37 through 56 of Test J1 are also items 1 through 20 of Test J2. The number of examinees used for the word comprehension scale construction varies between 412 sixth graders of elementary schools for Test A5 and 924 second graders of senior high schools for Test S1 (Shiba, 1978).

The model adopted for the item characteristic function of each vocabulary item is the logistic model which is given by (8.12), with $D = 1.7$, as the substitute for the normal ogive model. Note that Shiba did not use the three-parameter logistic model. This is based upon his belief that three-parameter models are not applicable for well-developed multiple-choice items, which he has formed through his many experiences in test construction and research.

The author found Shiba's research very interesting, especially in the following aspects.

(1) The word comprehension tests are very well constructed.

choosing each alternative carefully.

(2) Unlike many researchers in the United States, they have tried to make a full use of the distractors.

(3) Subjects were selected from many different age groups.

(X.2) Subjects Used in Shiba's Research

Each of the eleven tests was administered to a group of subjects who belong to a single school year, except for college students. Hereafter, for convenience, we shall use EL for elementary schools, JH for junior high schools, SH for senior high schools, and CS for colleges, and add the school year after each symbol. For instance, by SH2 we mean a group of subjects who are in the second year of senior high schools. The correspondence of the subject groups and the tests administered is summarized as follows:

A1 for EL1 (650), A2 for EL2 (650), A3 for EL3 (546),
A4 for EL4 (617), A5 for EL5 (599), A6 for EL6 (412),
J1 for JH1 (614), J2 for JH2 (758), S1 for SH1 (924),
S2 for SH2 (759), and U for CS (740).

where the numbers in parentheses indicate respective numbers of examinees. Note that JH3 and SH3 are not included in the data which are the basis of the word comprehension scale construction.

(X.3) Methods and Results of Shiba's Research

It is assumed that, for each of the eleven groups of examinees, the ability distribution is normal. The principal factor solution of factor analysis is applied for the tetrachoric correlation matrix for each group of examinees, using the largest absolute value of the correlation coefficient in each row, or column, as the communality. This step is also the process of validating the unidimensionality of ability. Figure 10-3-1 illustrates the resulting set of eigenvalues for Test J1 which was

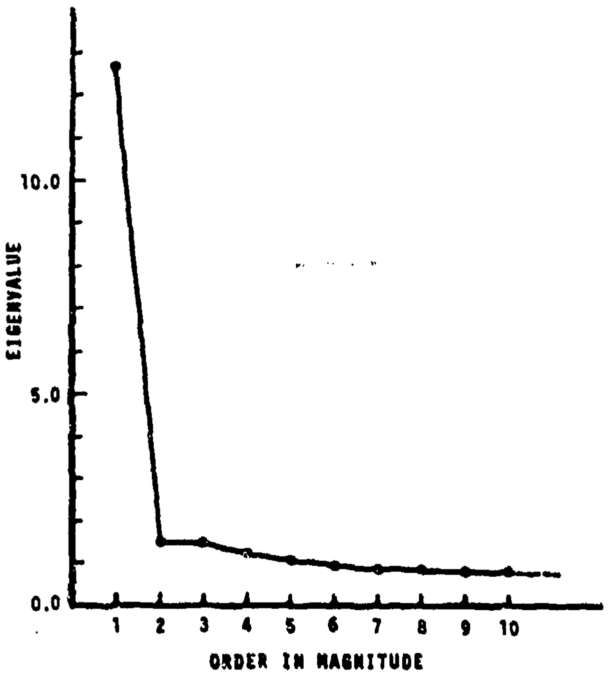


FIGURE 10-3-1

Eigenvalues of the Correlation Matrix of the Fifty-Five Items of Test J1, Ordered with Respect to Their Magnitudes. (Shiba's Data)

administered to 614 first year junior high school students. It turned out that the first eigenvalue is much larger than all the other eigenvalues, and thus the unidimensionality was confirmed. Hereafter, this first principal factor is treated as θ .

Let ρ_g be the factor loading (e.g., Lawley and Maxwell, 1971) of the first principal factor, or θ , for item g . The item discrimination parameter, a_g , is obtained by

$$(10.1) \qquad a_g = \rho_g(1-\rho_g)^{-1/2} \quad ,$$

Let $\Phi(u)$ denote the standard normal distribution function, such that

(10.2) $$\Phi(u) = (2\pi)^{-1/2}\int_{-\infty}^{u} e^{-t^2/2}\,dt \ .$$

The item difficulty parameter, b_g , is given by

(10.3) $$b_g = \Phi^{-1}(1-p_{gR})\ \rho_g^{\ -1} \ ,$$

where p_{gR} is the probability with which the examinee answers item g correctly. In practice, this is replaced by the frequency ratio, P_{gR} , to provide us with the estimate of b_g .

The eleven ability scales thus constructed are assumed to be on the same continuum, and they are integrated into a single scale. This equating is made through the ten subsets of items, each of which is shared by two adjacent tests. Let a_g and b_g be the item parameters estimated from the result of the first test, and a_g^* and b_g^* be those from the result of the second test. Denoting the two ability scales by θ and θ^* , respectively, we can write

(10.4) $$a_g(\theta-b_g) = a_g^*(\theta^*-b_g^*) \ ,$$

since the item characteristic functions, which follow the normal ogive model, of the same item g on the two ability scales must assume the same value for the corresponding values of θ and θ^* . Thus the functional relationship between θ and θ^* is given by

(10.5) $$\theta^* = (a_g/a_g^*)\theta + [b_g^*-(a_g/a_g^*)b_g] \ ,$$

which is linear, and the two coefficients are obtained from these four parameters. In practice, we obtain as many sets of coefficients as the number of common items, and we need to use some type of "average" of these coefficients for the scale transformation. Figure 10-3-2 presents the ability distributions of eleven subject groups after such transformations were made and the mean and the standard deviation of the distribution of J1 are taken as the

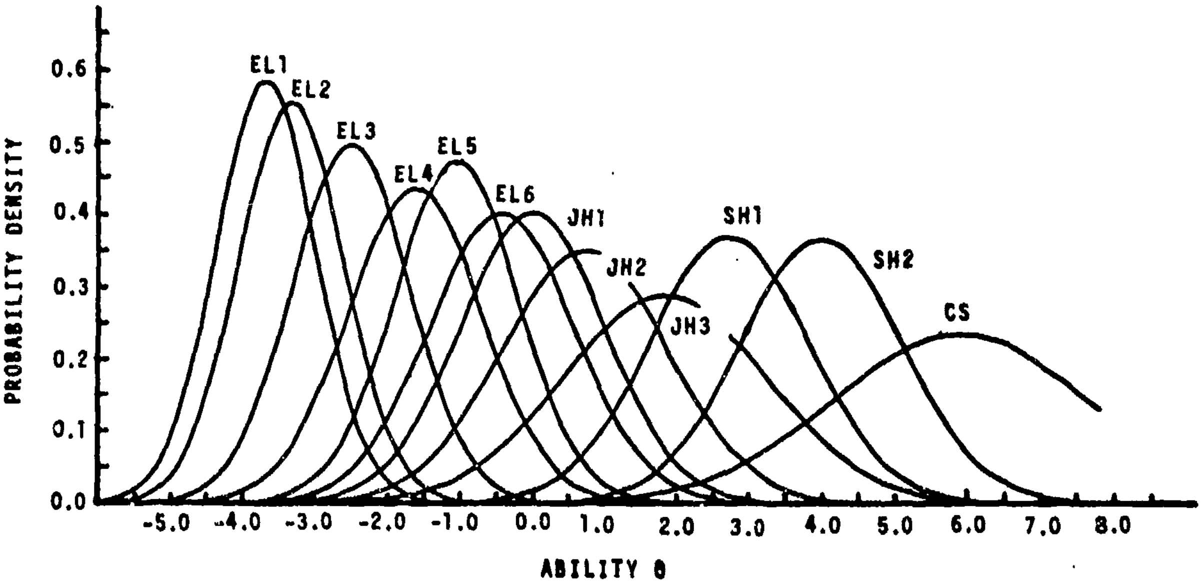


FIGURE 10-3-2

Estimated Density Functions of the Twelve Groups of Examinees, Which Are Assumed to Be Normal. The Ability Scale Is Defined in Such a Way that the Density Function of the Fisrt Grade Group of Junior High School (JH1) Is n(0,1) . (Shiba's Data)

origin and the unit for the new, integrated ability dimension.

The item characteristic function of each item on the new, integrated scale θ is approximated by the logistic function, which is given by (8.12). The maximum likelihood estimate, $\hat{\theta}_j$, of each examinee's ability is obtained through the equation

$$\sum_{g=1}^{n} a_g P_g(\hat{\theta}_j) = \sum_{g=1}^{n} a_g x_{gj} \tag{10.6}$$

(cf. Birnbaum, 1968), where x_{gj} is the binary item score of individual j for item g . The item information function of each test item, and then the test information of each test, are obtained (cf. Section III.4).

The theoretical frequency distribution of test score T for each test and examinee group can be written as

$$(10.7) \qquad N \sum_{V \varepsilon T} \sum_{u_g \varepsilon V} P_g(\theta)^{x_g} [1-P_g(\theta)]^{1-x_g} ,$$

where V is a response pattern of a vector of n items scores, and T is the test score given by

$$(10.8) \qquad T = \sum_{g=1}^{n} x_g .$$

This is used for the validation of the model and assumptions adopted in the process of analysis. The sample mean of the maximum likelihood estimates $\hat{\theta}$ of the subgroup of examinees, who selected each of the five alternatives is calculated, for each item of each test. A tailored test of the word comprehension is constructed by selecting an appropriate subset of items from these eleven tests, in such a way that an individual is directed to a next item which is chosen on the basis of the sample mean of $\hat{\theta}$ of the alternative he has selected for the present item.

The research conducted by Shiba and others includes more interesting data than were used in the word comprehension scale construction. Table 10-3-1 presents a part of them, in which the frequency distribution of the alternative selection by the first year students of junior high schools, and the mean of the maximum likelihood estimate of ability for each alternative are shown for nineteen items included in both Tests J1 and J2, and administered to four different subject groups, JH1, JH2(a), JH2(b) and JH3. In the same table, also presented is the discrepancy between the mean of $\hat{\theta}$ for the correct answer and the lowest mean $\hat{\theta}$ for one of the four wrong answers, under the heading, "largest discrepancy." The correct answers are always identified as the ones which have

TABLE 10-3-1

Mean of the Maximum Likelihood Estimates of Ability, $\hat{\theta}$, for Each of the Five Subgroups of Subjects Selecting Different Alternatives, for Each of the 19 Vocabulary Test Items, Together with the Actual Frequency Distributions (FRQ). The Difference between the Mean $\hat{\theta}$ of the Correct Subgroups and the Lowest Mean $\hat{\theta}$ Is Also Presented As Largest Discrepancy for Each Item. Test J1, Junior High School Grade 1

Item	Indices	Alternative					Total	Largest Discrepancy
		1	2	3	4	5		
37	Mean $\hat{\theta}$	0.401	-0.476	-0.482	-0.750	-0.148	572	1.151
	FRQ	287	50	59	59	117		
38	Mean $\hat{\theta}$							
	FRQ							
39	Mean $\hat{\theta}$	-0.192	-0.091	-0.270	-0.243	0.400	562	0.670
	FRQ	91	115	118	51	187		
40	Mean $\hat{\theta}$	0.071	-0.416	-0.336	0.310	-0.479	573	0.789
	FRQ	60	141	90	273	9		
41	Mean $\hat{\theta}$	-0.557	-1.007	-0.445	-0.456	0.254	573	1.261
	FRQ	53	20	23	85	392		
42	Mean $\hat{\theta}$	0.339	-0.570	0.036	-0.439	-0.387	570	0.909
	FRQ	247	21	121	84	97		
43	Mean $\hat{\theta}$	-0.512	0.376	-0.572	-0.245	-0.393	572	0.948
	FRQ	26	308	98	67	73		
44	Mean $\hat{\theta}$	-0.293	-0.547	-0.595	0.271	-0.318	569	0.866
	FRQ	119	67	14	333	36		
45	Mean $\hat{\theta}$	-0.638	-0.412	-0.636	0.395	-0.593	568	1.033
	FRQ	51	25	123	346	23		
46	Mean $\hat{\theta}$	0.444	-0.741	-0.325	-0.428	-0.534	568	1.185
	FRQ	296	46	44	164	18		
47	Mean $\hat{\theta}$	-0.261	0.270	-0.078	-0.426	-0.101	569	0.696
	FRQ	69	224	158	53	65		
48	Mean $\hat{\theta}$	-0.129	-0.024	-1.013	-0.467	0.412	564	1.425
	FRQ	81	100	58	67	258		
49	Mean $\hat{\theta}$	-0.339	-0.390	-0.284	-0.464	0.309	573	0.773
	FRQ	115	31	42	70	315		
50	Mean $\hat{\theta}$	0.349	-0.256	-1.015	-0.317	-0.385	571	1.364
	FRQ	308	46	35	86	96		
51	Mean $\hat{\theta}$	-0.137	-0.640	-0.077	-0.136	0.429	560	1.069
	FRQ	89	82	75	113	201		
52	Mean $\hat{\theta}$	-0.219	0.291	-0.110	-0.608	-0.095	565	0.899
	FRQ	116	235	80	34	100		
53	Mean $\hat{\theta}$	-0.071	-0.030	-0.453	0.527	-0.241	572	0.980
	FRQ	163	51	34	143	181		
54	Mean $\hat{\theta}$	0.132	-0.060	-0.084	-0.037	-0.283	561	0.415
	FRQ	182	111	100	142	26		
55	Mean $\hat{\theta}$	0.114	-0.278	-0.172	-0.533	0.690	571	1.223
	FRQ	27	72	317	29	126		
56	Mean $\hat{\theta}$	-0.460	-0.113	-0.412	0.742	0.015	572	1.202
	FRQ	104	101	115	141	111		

the highest means of $\hat{\theta}$.

(X.4) Distractors As Resources of Information

Shiba's research is based upon his belief in the usefulness of distractors as important resources of information, in addition to the correct answer. This is the same belief which the author has kept in mind for many years (cf. Samejima, 1968). As far as we score the multiple-choice test item correct or incorrect and treat it as a binary item, it can never surpass the free-response test item, but will always stay as a "blurred" image of the free-response test item, owing to the noise caused by the examinee's guessing behavior, etc. If we make the full use of the information given by distractors, however, then the multiple-choice test item will have the merit of its own, and can even be more informative than the free-response test item.

It is researchers' responsibility to increase the efficiency in mental measurement. To ignore whatever legitimate information we can obtain from our research data is against this principle. If distractors can serve for this purpose, we should certainly not to stay with models like the three-parameter logistic model, in which all the wrong answers given as alternatives in the multiple-choice test item are treated as being equivalent, without any information of their own. It will be worth our effort to investigate Informative Distribution Model rather than to stay with the Equivalent Distractor Model (cf. Sections IX.9 and IX.10).

(X.5) Mathematical Models in Physics and in Psychology

The role of mathematical models in any science may be to describe its reality following an appropriate rationale. We must recognize, however, some difference between the role of mathematical models in pure natural sciences, like physics, and that in psychology. This difference comes from the fact that, while in physics it is impossible or meaningless to change natural phenomena to which objects react, in psychology many phenomena to which persons react are also made by persons, and it is quite legitimate to change them for good causes.

The latter logic is directly applicable to models for the multiple-choice test item. An important implication is that we may be able to do better than supplying mathematical models for the existing test items rather passively. If we conceive of some mathematical models which, in theory, will enhance the efficiency in mental measurement, we shall be able to advise test constructors to develop the types of multiple-choice test items which follow our models, instead of accepting whatever test items they produce. We can also adjust the pressure and its directions which are put upon examinees, by changing our instructions appropriately. To give an example, we can effectively discourage our examinees to guess, or to skip items.

(X.6) Normal Ogive Model on the Graded Response Level and Bock's Multinomial Model

Normal ogive model, which was originally introduced as a model for a binary, free-response test item, has been expanded to fit a more general case, in which an item is graded into more than two item score categories (Samejima, 1969, 1972). Bock has proposed a multinomial model (Bock, 1972), for the multiple-choice test item. It has been pointed out (Samejima, 1972) that, although Bock's model was originally developed for nominal categories, i.e., the categories which are not ordered among themselves, it can be considered as a model in the heterogeneous case of the graded response level.

Let g be a multiple-choice item, h , i or k be one of its m_g alternatives, and X_{hg} , X_{ig} or X_{kg} be the response tendency for the alternative, h , i or k . When any two alternatives, h and k , are compared alone, the probability with which h is chosen in preference to k is assumed to be a function of ability θ , and is denoted by $\pi_{hk}(\theta;g)$. Thus we can write

(10.9) $$\pi_{hk}(\theta;g) + \pi_{kh}(\theta;g) = 1 \quad .$$

When the comparison is made among m_g ($\geq$2) alternatives, the conditional probability with which the alternative h is chosen in preference to all the other (m_g-1) alternatives, given θ , is denoted by $P_h(\theta;g)$, and we have

(10.10) $$\sum_{h=1}^{m_g} P_h(\theta;g) = 1 \quad .$$

We shall define a variable $X_{hk;g}$, such that

(10.11) $$X_{hk;g} = X_{hg} - X_{kg} \quad ,$$

i.e., the difference between the two response tendencies, X_{hg} and X_{kg} .

Hereafter, for simplicity, we shall drop the subscript g , whenever it is clear that we are dealing with only one multiple-choice item. Thus, in such a case, $\pi_{hk}(\theta)$ is used for $\pi_{hk}(\theta;g)$, X_{hk} for $X_{hk;g}$, and so forth.

In the multinomial model, it is assumed that; 1) the conditional distribution of X_k , given θ , is normal, with $\mu_k(\theta;g)$, or $\mu_k(\theta)$, and $\sigma_k(\theta;g)$, or $\sigma_k(\theta)$, as the two parameters; 2) X_k's are conditionally, mutually independent, given θ ; and 3) the ratio of the probabilities with which the two alternatives are chosen, respectively, is invariant for the set of alternatives among which the two alternatives are compared. Thus for the third assumption we can write

(10.12) $$P_h(\theta)/P_k(\theta) = \pi_{hk}(\theta)/\pi_{kh}(\theta) \quad .$$

From the first two of these assumptions, it is derived that the conditional distribution of X_{hk} , given θ , is also normal, with $\mu_{hk}(\theta;g)$, or $\mu_{hk}(\theta)$, and $\sigma_{hk}(\theta;g)$, or $\sigma_{hk}(\theta)$, as the two parameters, which are given by

$$(10.13) \qquad \mu_{hk}(\theta) = \mu_h(\theta) - \mu_k(\theta) \quad ,$$

and

$$(10.14) \qquad \sigma_{hk}(\theta) = [\sigma_h^2(\theta) + \sigma_k^2(\theta)]^{1/2} \quad .$$

We can also write for $\pi_{hk}(\theta)$ and $\pi_{kh}(\theta)$ such that

$$(10.15) \qquad \pi_{hk}(\theta) = (2\pi)^{-1/2}\sigma_{hk}(\theta)^{-1}\int_0^\infty \exp[-\{X_{hk}-\mu_{hk}(\theta)\}^2/\{2\sigma_{hk}^2(\theta)\}]dX_{hk} \quad ,$$

and

$$(10.16) \qquad \pi_{kh}(\theta) = (2\pi)^{-1/2}\sigma_{hk}(\theta)^{-1}\int_{-\infty}^0 \exp[-\{X_{hk}-\mu_{hk}(\theta)\}^2/\{2\sigma_{hk}^2(\theta)\}]dX_{hk} \quad .$$

Now we shall use the logistic approximation to the normal distribution function, which is, with $D = 1.7$, given by

$$(10.17) \qquad (2\pi)^{-1/2}\int_{-\infty}^u e^{-u^2/2}\,du \doteq [1+\exp\{-Du\}]^{-1} \quad .$$

Thus we obtain from (10.13), (10.14), (10.15), (10.16) and (10.17)

$$\begin{aligned}(10.18) \qquad \pi_{hk}(\theta)/\pi_{kh}(\theta) &\doteq [1+\exp\{D\mu_{hk}(\theta)/\sigma_{hk}(\theta)\}] \\ &\qquad [1-\{1+\exp[D\mu_{hk}(\theta)/\sigma_{hk}(\theta)]\}^{-1}] \\ &= \exp[D\mu_{hk}(\theta)/\sigma_{hk}(\theta)] \\ &= \exp[D\{\mu_h(\theta)-\mu_k(\theta)\}/\{\sigma_h^2(\theta)+\sigma_k^2(\theta)\}^{1/2}] \quad .\end{aligned}$$

From (10.12) and (10.18) we can write

$$(10.19)\qquad [P_k(\theta)]^{-1} = \sum_{i=1}^{m}[P_i(\theta)/P_k(\theta)] = \sum_{i=1}^{m}[\pi_{ik}(\theta)/\pi_{ki}(\theta)]$$
$$\doteq \sum_{i=1}^{m} \exp[D\{\mu_i(\theta)-\mu_k(\theta)\}/\{\sigma_i^2(\theta)+\sigma_k^2(\theta)\}^{1/2}] \quad ,$$

and then

$$(10.20)\qquad P_h(\theta) = P_k(\theta)[\pi_{hk}(\theta)/\pi_{kh}(\theta)]$$
$$\doteq \exp[D\{\mu_h(\theta)-\mu_k(\theta)\}/\{\sigma_h^2(\theta)+\sigma_k^2(\theta)\}^{1/2}]$$
$$[\sum_{i=1}^{m} \exp\{D[\mu_i(\theta)-\mu_k(\theta)]/[\sigma_i^2(\theta)+\sigma_k^2(\theta)]^{1/2}\}]^{-1} \quad ,$$

for $h=1,2,\ldots,m$. Note that k is an arbitrarily chosen, fixed alternative.

If we add two other assumptions such that: 4) the regression of the response tendency X_h $(h=1,2,\ldots,m)$ is linear; and 5) the conditional variance of X_h , given θ , is constant, i.e.,

$$(10.21)\qquad \mu_h(\theta) = a_h^*\theta + c_h^*$$

and

$$(10.22)\qquad \sigma_h^2(\theta) = \sigma_h^2 \quad ,$$

then we can write

$$(10.23)\qquad D[\mu_h(\theta)-\mu_k(\theta)]/[\sigma_h^2(\theta)+\sigma_k^2(\theta)]^{1/2} = a_h\theta + c_h \quad ,$$

where

$$(10.24)\qquad a_h = D(a_h^*-a_k^*)/(\sigma_h^2+\sigma_k^2)^{1/2}$$

and

$$(10.25) \qquad c_h = D(c^*_h - c^*_k)/(\sigma_h^2+\sigma_k^2)^{1/2} \quad .$$

Substituting (10.23) into (10.20), we obtain

$$(10.26) \qquad P_h(\theta) = \exp[a_h\theta+c_h][\sum_{i=1}^{m} \exp\{a_i\theta+c_i\}]^{-1} \quad .$$

Thus (10.26) specifies the operating characteristic of the category h in the multinomial model. Note that both a_i's and c_i's in (10.26) are of *arbitrary origins*, for we have for arbitrary d and e

$$\begin{aligned}(10.27) \qquad P_h(\theta) &= \exp[a_h\theta+c_h]\exp[d\theta+e][\sum_{i=1}^{m} \exp\{d\theta+e\}\exp\{a_i\theta+c_i\}]^{-1} \\ &= \exp[(a_h+d)\theta+(c_h+e)][\sum_{i=1}^{m} \exp\{(a_i+d)\theta+(c_i+e)\}]^{-1} \quad .\end{aligned}$$

While in the multinomial model we assume m different response tendencies and their conditional independence, and the invariance of the ratio of the two probabilities of alternative selection, in the normal ogive model on the graded response level, we assume that there exists a *single response tendency*, or item variable, X_g , or X , behind the selection of any one of the m alternatives, and the conditional distribution of X , given θ , is normal, with $\mu(\theta)$ and $\sigma(\theta)$ as the two parameters. In addition to this first assumption, we also assume for the normal ogive model that: 2) the whole dimension of the item variable X is divided into m subintervals; and 3) the alternative h will be selected if the examinee's response tendency is in the subinterval assigned to that category. We can write

$$(10.28) \qquad P_h(\theta) = [2\pi]^{-1/2}[\sigma(\theta)]^{-1}\int_{\gamma_{h-1}}^{\gamma_h} \exp[-\{u-\mu(\theta)\}^2/\{2\sigma(\theta)^2\}]du \quad ,$$

where γ_h is the upper endpoint of the subinterval of X which is assigned to the category h , and we have

$$(10.29) \qquad \begin{cases} \gamma_0 = -\infty \\ \gamma_m = \infty \quad . \end{cases}$$

The additional two assumptions, 4) and 5), for the multinomial model, which are formulated by (10.21) and (10.22), respectively, are also adopted for the item variable X in the normal ogive model. Thus we can write for the conditional expectation, or regression, of X on θ and the conditional variance of X, given θ,

$$\mu(\theta) = a^*\theta + c^* \quad , \tag{10.30}$$

and

$$\sigma^2(\theta) = \sigma^2 \quad . \tag{10.31}$$

Substituting (10.30) and (10.31) into (10.28), we obtain

$$\begin{aligned} P_h(\theta) &= [2\pi]^{-1/2}\sigma^{-1}\int_{\gamma_{h-1}}^{\gamma_h} \exp[-(u-a^*\theta-c^*)^2/(2\sigma^2)]\,du \\ &= [2\pi]^{-1/2} \int_{(\gamma_{h-1}-a^*\theta-c^*)/\sigma}^{(\gamma_h-a^*\theta-c^*)/\sigma} \exp[-t^2/2]\,dt \\ &= [2\pi]^{-1/2} \int_{(a^*\theta+c^*-\gamma_h)/\sigma}^{(a^*\theta+c^*-\gamma_{h-1})/\sigma} \exp[-t^2/2]\,dt \quad , \end{aligned} \tag{10.32}$$

where

$$t = (u-a^*\theta-c^*)/\sigma \quad . \tag{10.33}$$

We define the item parameters, a_g, or a, and b_{hg}, or b_h, such that

$$a = a^*/\sigma \tag{10.34}$$

and

$$b_h = (\gamma_h-c^*)/a^* \quad , \tag{10.35}$$

where

$$(10.36) \qquad \begin{cases} b_0 = -\infty \\ b_m = \infty \end{cases} .$$

Substituting (10.34) and (10.35) into (10.36), we obtain for the normal ogive model on the graded response level,

$$(10.37) \qquad P_h(\theta) = [2\pi]^{-1/2} \int_{a(\theta-b_h)}^{a(\theta-b_{h-1})} \exp[-t^2/2]\ dt \quad .$$

We have seen in the preceding paragraphs that in both the normal ogive model on the graded response level and Bock's multinomial model the normal assumption is made for the conditional distribution of the response tendency, given ability θ . The biggest difference between the two models is that, in the normal ogive model, a single item variable is assumed behind the examinee's selection behavior, whereas in the multinomial model a separate response tendency is assumed for each of the m alternatives.

These two models, and logistic model on the graded response levels (Samejima, 1969, 1972), whose operating characteristic, $P_h(\theta)$ $(h=1,2,...,m)$, is given by

$$(10.38) \qquad P_h(\theta) = [1-\exp\{-Da(b_h-b_{h-1})\}][1+\exp\{-Da(\theta-b_{h-1})\}]^{-1} [1+\exp\{Da(\theta-b_h)\}]^{-1} \quad ,$$

can be used as models for the multiple-choice test item, in such a testing situation that guessing is extremely discouraged by our instructions. "No Answers" will be treated as the response of the lowest rank in such a situation.

(X.7) A New Family of Models for the Multiple-Choice Test Items

Suppose that our multiple-choice test item is constructed well enough to provide us with $(m-1)$ distractors, which have certain levels of plausibility to attract examinees. Suppose, further, that there is some simple statistical relationship between each distractor and ability θ, i.e., the conditional probability with which the examinee chooses the distractor h as the correct answer in comparison with all the other (m-1) alternatives, given θ, increases in θ up to a certain level of θ, and then decreases in θ. This implies that there may be individuals who do not know the answer, nor recognize the plausibility of any distractor. Suppose that the conditional probability with which the examinee belongs to this category, given θ, is strictly decreasing in ability θ. If the item characteristic function is strictly increasing in ability θ with zero and unity as its two asymptotes, and if, in addition, the two asymptotes of the "plausibility" function for each of the (m-1) distractors are uniformly zero, and those of the conditional probability for the "no recognition" category are unity and zero, respectively, then both normal ogive model on the graded response level and the multinomial model are included in this type of models.

This type of models is suitable only if the supervision is strict and the examinees are extremely discouraged to guess when they do not know the right answer. It may be more realistic to assume, however, that in most testing situations the pressure for success is so strong that the examinees do guess when they have no idea about the correct answer. Suppose that these examinees guess randomly, and select one of the m alternatives with equal probability. Thus we obtain a new family of models, which includes modified forms of such models as normal ogive and logistic models on the graded response level and the multinomial model.

Let $P_{x_g}(\theta)$ be the operating characteristic of the graded response category x_g $(=0,1,2,...,m_g)$, whose mathematical form is given as $P_h(\theta;g)$ in (10.26), (10.37) of (10.38), or of any other model of similar characteristics. For convenience, we shall call these models as models of Type I on the graded response level. To be specific, models of Type I are those which satisfy the following.

(1) $P_{x_g}(\theta)$ is strictly decreasing in θ , with unity and zero as its two asymptotes, for $x_g = 0$.

(2) $P_{x_g}(\theta)$ is unimodal with zero as its two asymptotes, for $x_g = 1, 2, \ldots, (m_g - 1)$.

(3) $P_{x_g}(\theta)$ is strictly increasing in θ , with zero and unity as its two asymptotes, for $x_g = m_g$.

The above conditions for Type I models also imply that $\sum_{s=x_g}^{m_g} P_s(\theta)$ is strictly increasing in θ with zero and unity as its two asymptotes, for $x_g = 1, 2, \ldots, m_g$.

We use this additional response category $x_g = 0$ for those who have no idea at all as to which alternative the correct answer is. Thus the probability with which the examinee belongs to this category is strictly decreasing in θ , with unity and zero as its two asymptotes. We assume that the $(m_g - 1)$ distractors of the multiple-choice item g have an implicit order among themselves. Thus the response categories, $x_g = 1, 2, \ldots, (m_g - 1)$, are used for the $(m_g - 1)$ distractors, and their operating characteristics of the distractors are unimodal, with zero as their two asymptotes, respectively. The other category, $x_g = m_g$, is for the correct answer, and its operating characteristic is strictly increasing in θ with zero and unity as its two asymptotes. Since, in reality, the examinees who belong to the category, $x_g = 0$, are assumed to guess randomly, however, the operating characteristic for this response category disappears, and those of the other categories, or the m_g alternatives, are affected by this random guessing. The operating characteristic of the alternative h can be written, therefore, such that

$$(10.39) \qquad P_h(\theta;g) = P_{x_g}(\theta;x_g = h) + (1/m_g)P_{x_g}(\theta;x_g = 0) \ .$$

Thus we have obtained a new family of models for the multiple-choice item. When $P_{x_g}(\theta)$ follows the normal ogive model on the graded response level, $P_h(\theta;g)$, or $P_h(\theta)$, takes on a form such that

$$(10.40) \qquad P_h(\theta) = (2\pi)^{-1/2}[\int_{a(\theta-b_{h+1})}^{a(\theta-b_h)} e^{-u^2/2}du + (1/m)\int_{a(\theta-b_1)}^{\infty} e^{-u^2/2}du] ,$$

where $a > 0$, and

$$(10.41) \qquad -\infty < b_1 < b_2 < \ldots < b_m < b_{m+1} = \infty .$$

For simplicity, we shall call it Model A of Type I for the multiple-choice item. When $P_{x_g}(\theta)$ in (10.39) is specified by the logistic model on the graded response level, we can write

$$(10.42) \qquad P_h(\theta) = [1-\exp\{-Da(b_{h+1}-b_h)\}][1+\exp\{-Da(\theta-b_h)\}]^{-1}$$
$$[1+\exp\{Da(\theta-b_{h+1})\}]^{-1} + [m\{1+\exp[Da(\theta-b_1)]\}]^{-1} ,$$

where $a > 0$, and the inequality (10.41) also holds. We shall call it Model B of Type I for the multiple-choice item. When the operating characteristic of the category x_g in the multinomial model is substituted for $P_{x_g}(\theta)$ in (10.39), we obtain

$$(10.43) \qquad P_h(\theta) = [\exp\{a_h\theta+c_h\} + (1/m)\exp\{a_0\theta+c_0\}][\sum_{i=0}^{m} \exp\{a_i\theta+c_i\}]^{-1} ,$$

where

$$(10.44) \qquad a_0 < a_1 < a_2 < \ldots < a_m .$$

We shall call it Model C of Type I for the multiple-choice item, or Bock-Samejima model for the multiple-choice item.

For the purpose of illustration, Figure 10-7-1 presents the operating characteristics of the six response categories, following the normal ogive model on the graded response level, with $a_g = 2.00$, $b_1 = -2.00$, $b_2 = -1.00$, $b_3 = 0.00$, $b_4 = 1.00$ and $b_5 = 2.00$. The

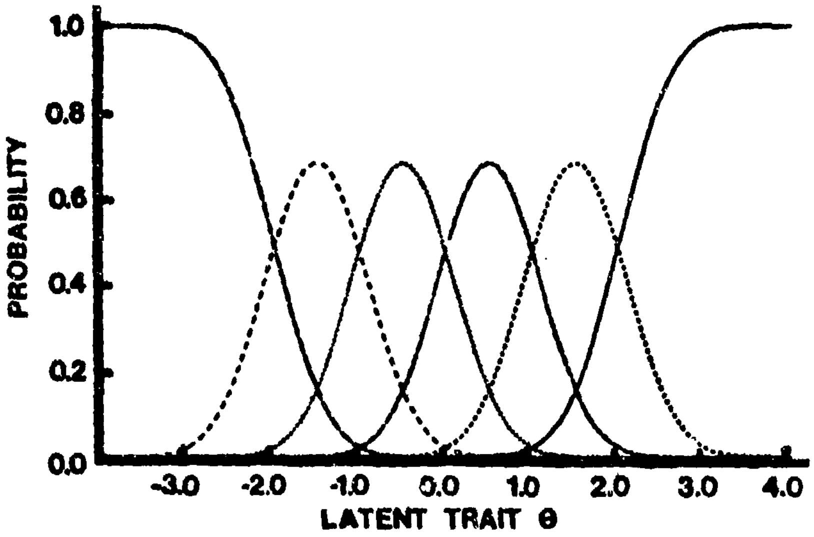


FIGURE 10-7-1

Operating Characteristics of Six Item Response Categories Following the Normal Ogive Model, with $a_g = 2.00$, $b_1 = -2.00$, $b_2 = -1.00$, $b_3 = 0.00$, $b_4 = 1.00$ and $b_5 = 2.00$.

modal point of the operating characteristic of each of the $(m-1)$ intermediate categories is given by $(b_h + b_{h+1})/2$ (Samejima, 1969). Figure 10-7-2 presents the corresponding operating characteristics of the five alternatives following Model A of Type I for the multiple-choice item. We can see that these curves are no longer symmetric for $h = 1,2,3$ and 4 . It is indicated in the figure that the asymptotes of these operating characteristics at $\theta = -\infty$ are uniformly 0.2 , or $1/m$.

Figure 10-7-3 and 10-7-4 present the operating characteristics of the six response categories in the logistic model, and those of the five alternatives in Model B, which follow the mathematical form given as (10.42), for a hypothetical multiple-choice item. In these figures, item parameters are : $a_1 = 1.00$, $b_1 = -1.50$, $b_2 = -1.00$, $b_3 = -0.50$, $b_4 = 0.00$ and $b_5 = 0.50$.

Figures 10-7-5 and 10-7-6 present the operating characteristics of the five response categories following the multinomial model, and

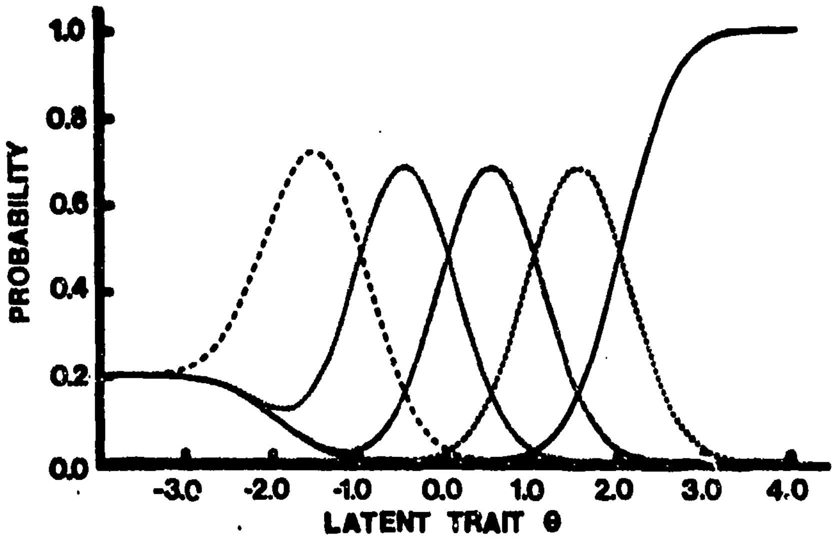


FIGURE 10-7-2

Operating Characteristics of Five Alternatives Following the Model A of Type I for the Multiple-Choice Item. The Parameters Are: $a_g = 2.00$, $b_1 = -2.00$, $b_2 = -1.00$, $b_3 = 0.00$, $b_4 = 1.00$ and $b_5 = 2.00$.

those of the four alternatives in Model C, which are given by (10.43), as the third example. The item parameters for this pair of operating characteristics are : $a_1 = -1.00$, $a_2 = -0.50$, $a_3 = 0.00$, $a_4 = 0.50$, $a_5 = 1.00$, $c_1 = 1.00$, $c_2 = -0.50$, $c_3 = 0.00$, $c_4 = -1.25$ and $c_5 = 0.75$.

(X.8) Basic Functions and Information Functions of the New Models

The basic function, $A_{x_g}(\theta)$, which is defined by (3.14) in Section III.5., has an essential role in the numerical solution of the maximum likelihood estimation of the examinee's ability. A sufficient condition that a model defined on the graded response level provides us with a unique maximum likelihood estimate for every possible response pattern, or the unique maximum condition, is that the basic function is strictly decreasing in θ with a non-negative asymptote at $\theta \to -\infty$ and a non-positive asymptote as $\theta \to \infty$, with respect to every item response category (cf. Samejima, 1969, 1972).

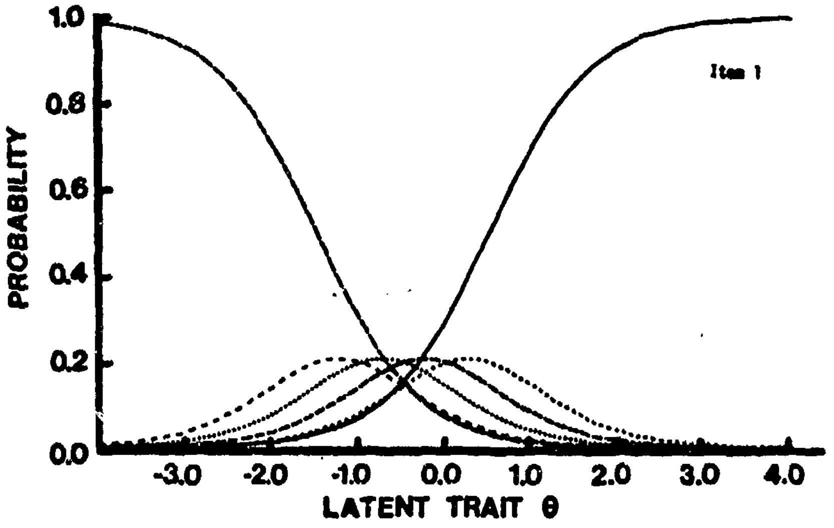


FIGURE 10-7-3

Operating Characteristics of Six Item Response Categories Following the Logistic Model, with a_g = 1.00 , b_1 = -1.50 , b_2 = -1.00 , b_3 = -0.50 , b_4 = 0.00 and b_5 = 0.50 .

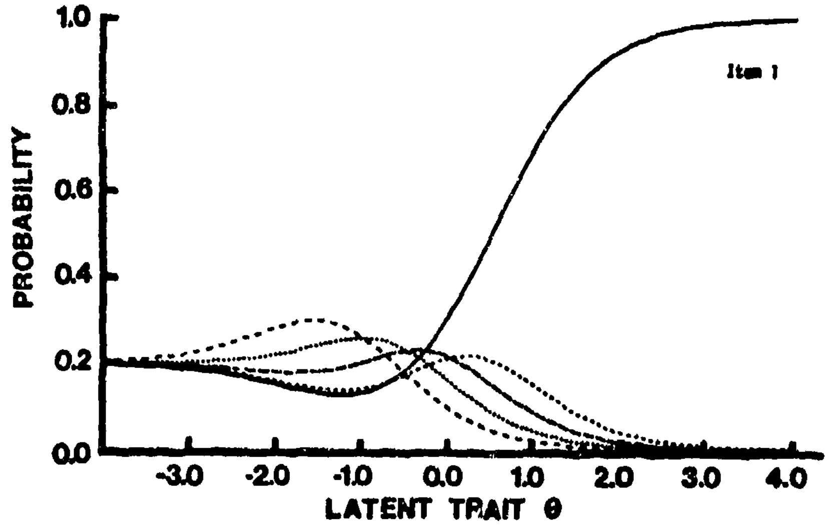


FIGURE 10-7-4

Operating Characteristics of Five Alternatives Following Model B of Type I for the Multiple-Choice Item, with the Parameters, a_g = 1.00 , b_1 = -1.50 , b_2 = -1.00 , b_3 = -0.50 , b_4 = 0.00 and b_5 = 0.50 .

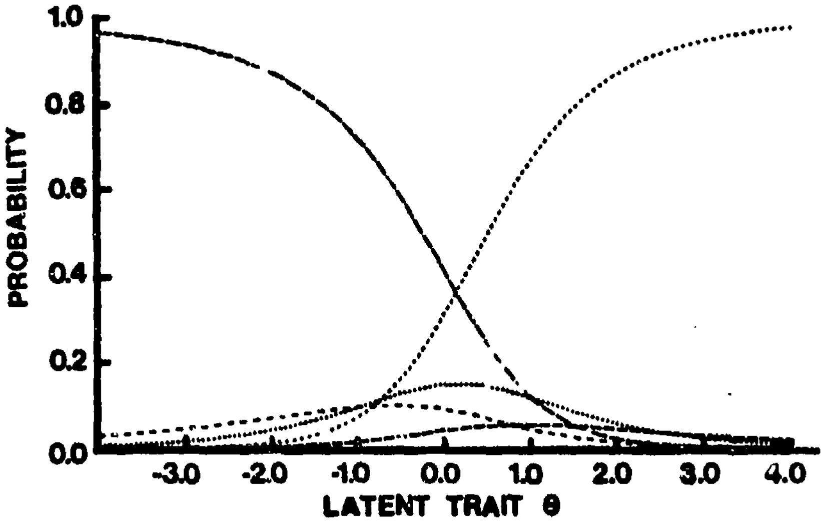


FIGURE 10-7-5

Operating Characteristics of Five Item Response Categories Following the Multinomial Model. The Item Parameters Are: $a_1 = -1.000$, $a_2 = -0.500$.
$a_3 = 0.000$. $a_4 = 0.500$. $a_5 = 1.000$; $c_1 = 1.000$. $c_2 = -0.500$,
$c_3 = 0.000$, $c_4 = -1.250$, $c_5 = 0.750$.

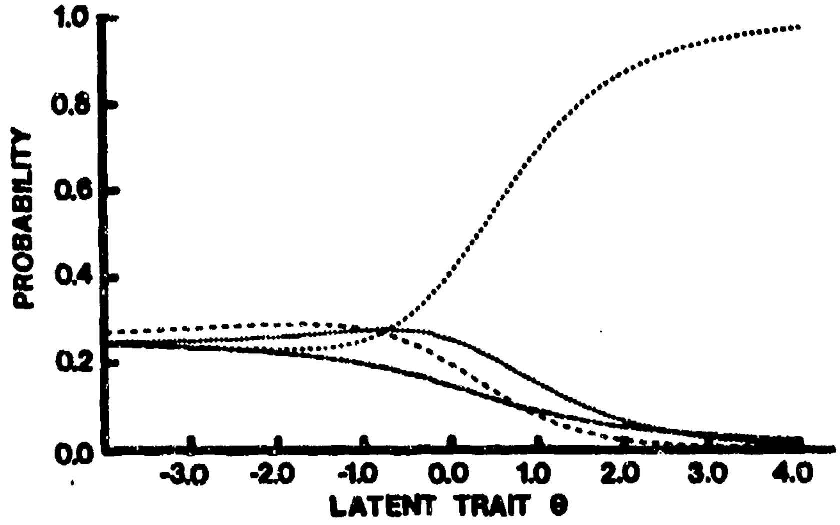


FIGURE 10-7-6

Operating Characteristics of Four Alternatives Following Model C of Type I for the Multiple-Choice Item, with the Parameters, $a_1 = -1.000$, $a_2 = -0.5000$,
$a_3 = 0.000$, $a_4 = 0.500$, $a_5 = 1.000$; $c_1 = 1.000$, $c_2 = -0.500$,
$c_3 = 0.000$, $c_4 = -1.250$, $c_5 = 0.750$.

The analogous basic function can be defined for the multiple-choice item. We have for the basic function, $A_h(\theta)$, of the alternative h

$$(10.45) \qquad A_h(\theta) = \frac{\partial}{\partial\theta} \log P_h(\theta) = \frac{\partial}{\partial\theta} P_h(\theta) \, [P_h(\theta)]^{-1} \ .$$

For the alternative information function, $I_h(\theta)$, we can write

$$(10.46) \qquad I_h(\theta) = -\frac{\partial^2}{\partial\theta^2} \log P_h(\theta) = [A_h(\theta)]^2 - \frac{\partial^2}{\partial\theta^2} P_h(\theta) [P_h(\theta)]^{-1} \ .$$

The item information function of the multiple-choice item is the conditional expectation of the alternative information function, given θ, such that

$$(10.47) \qquad I_g(\theta) = \sum_{h=1}^{m} I_h(\theta) \, P_h(\theta) = \sum_{h=1}^{m} [A_h(\theta)]^2 \, P_h(\theta) \ .$$

It should be noted that these basic functions and information functions assume more complicated forms than the corresponding functions for the graded item response categories, if we adopt one of the models for the multiple-choice item, i.e., Models A, B and C. We shall take Model B of Type I as an example, and observe its basic functions and information functions, which are given by (10.45), (10.46) and (10.47). Comparison will be made between these functions in Model B and those in the logistic model on the graded response level, which share the same parameters.

Let $P_h^*(\theta)$ be such that

$$(10.48) \qquad P_h^*(\theta) = [1 + \exp\{-Da(\theta - b_h)\}]^{-1} \ .$$

Then we can rewrite (10.42) for the operating characteristic of the alternative h in the form

$$(10.49) \qquad P_h(\theta) = [1-\exp\{-Da(b_{h+1}-b_h)\}]\, P^*_h(\theta)\, [1-P^*_{h+1}(\theta)] + (1/m)[1-P^*_1(\theta)] \ .$$

From (10.49) we obtain the first and second derivatives of $P_h(\theta)$ such that

$$(10.50) \qquad \frac{\partial}{\partial\theta} P_h(\theta) = Da[\{1-\exp[-Da(b_{h+1}-b_h)]\}P^*_h(\theta)\{1-P^*_{h+1}(\theta)\} \{1-P^*_h(\theta)-P^*_{h+1}(\theta)\}] - (1/m)DaP^*_1(\theta)[1-P^*_1(\theta)]$$

and

$$(10.51) \qquad \frac{\partial^2}{\partial\theta^2} P_h(\theta) = D^2a^2[\{1-\exp[-Da(b_{h+1}-b_h)]\}\{[1-P^*_h(\theta)-P^*_{h+1}(\theta)]^2 - P^*_h(\theta)[1-P^*_h(\theta)] - P^*_{h+1}(\theta)[1-P^*_{h+1}(\theta)]\}] - (1/m)D^2a^2P^*_1(\theta)[1-P^*_1(\theta)][1-2P^*_1(\theta)] \ .$$

It is noted that the last term in each of (10.49), (10.50) and (10.51) is the term which makes the function different from the corresponding function in the logistic model on the graded response level. The amount of effect caused by these additional terms on the basic functions and the information functions for different levels of θ depends upon the parameter b_1 , or b_h for $h = 1$. If these additional terms do not exist, i.e., in the logistic model on the graded response level, we can write for the basic functions and the information functions

$$(10.52) \qquad A_h(\theta) = Da[1-P^*_h(\theta)-P^*_{h+1}(\theta)] \ ,$$

$$(10.53) \qquad I_h(\theta) = D^2a^2[P^*_h(\theta)\{1-P^*_h(\theta)\} + P^*_{h+1}(\theta)\{1-P^*_{h+1}(\theta)\}] \ ,$$

where $h = 0,1,2,\ldots,m$, and

$$(10.54) \qquad I_g(\theta) = D^2a^2 \sum_{h=0}^{m} [1-P^*_h(\theta)-P^*_{h+1}(\theta)]^2[P^*_h(\theta)-P^*_{h+1}(\theta)] \ .$$

Figure 10-8-1 presents the basic functions of the six categories in the logistic model, which are given by (10.52), and those of the five alternatives in Model B of Type I for the multiple-choice item, which were obtained by substituting (10.49) and (10.50) into (10.45), for the hypothetical test item, whose operating characteristics are

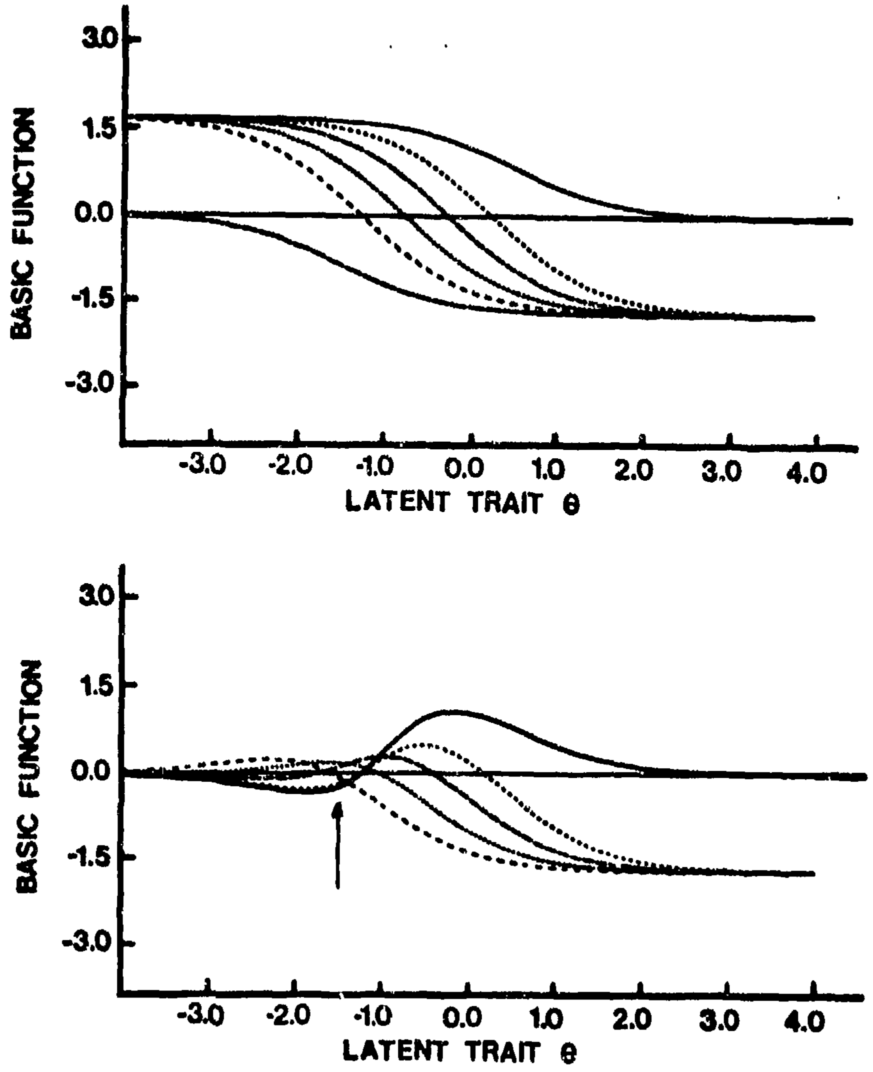


FIGURE 10-8-1

Basic Functions of Six Item Response Categories in the Logistic Model (Above), and Those of Five Alternatives in Model B (Below). The Item Parameters Are: $a_g = 1.00$, $b_1 = -1.50$, $b_2 = -1.00$, $b_3 = -0.50$, $b_4 = 0.00$ and $b_5 = 0.50$.

shown in Figures 10-7-3 and 10-7-4. As we can see in the first graph, all the six basic functions in the logistic model are strictly decreasing in θ, with the common asymptote, $1.7a$, at $\theta \to -\infty$ for $h = 2,3,4,5,6$ and $-1.7a$ at $\theta \to \infty$ for $h = 1,2,3,4,5$, while for $h = 6$ the asymptote at $\theta \to \infty$ is zero and for $h = 1$ the one at $\theta \to -\infty$ is zero, respectively (cf. Samejima, 1969). It should also be noted that for the four intermediate categories, $h = 2,3,4,5$, the basic functions take on zero at $\theta = (b_h+b_{h+1})/2$.

We find quite a contrasting set of five basic functions in the second graph of Figure 10-8-1. In fact, none of these basic functions are strictly decreasing in θ, but each has a unique modal point, and, except for $h = 1$ a unique local minimum also. The common asymptote at $\theta \to \infty$ for the alternatives excluding the correct answer is $-1.7a$, just as in the logistic model, and the other common asymptote at $\theta \to -\infty$, along with the asymptote at $\theta \to \infty$ for the correct answer, is zero, as is expected from (10.49) and (10.50). It is very obvious from these results that Model B does not satisfy the unique maximum condition, and, therefore, a unique maximum likelihood estimate is not assured for every possible response pattern. We need to pursue the characteriestics of this model further and find out some practical solution for this problem, therefore, as was done for the three-parameter logistic model (Samejima, 1973).

We notice that these basic functions are practically identical with the corresponding curves in the logistic model, for certain intervals of higher ability. Needless to say, it is desirable if these intervals start from relatively lower levels of ability θ. It is obvious that the lower endpoint of such an interval depends upon the parameter b_1, which is indicated by an arrow in the graph of Model B.

Figure 10-8-2 presents the alternative information functions in the logistic model, and the corresponding alternative information functions in Model B, in the upper and lower parts, respectively, of the same multiple choice item. Among each of the two sets of six and five

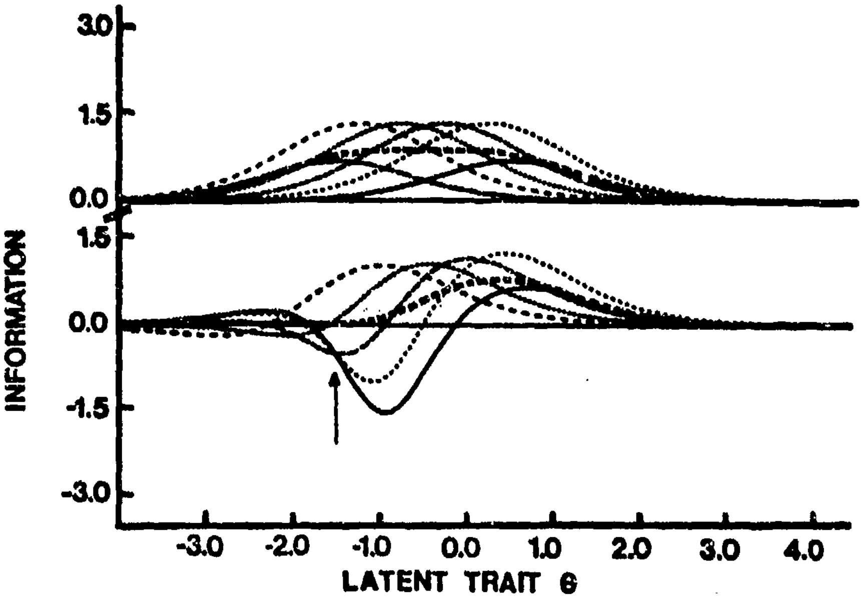


FIGURE 10-8-2

Item Response Information Functions (Various Thinner Curves) and the Item Information Function (Heavy Dashes) in the Logistic Model (Above) and Those in Model B of Type I for the Multiple-Choice Item (Below). The Item Parameters Are: $a_g = 1.00$, $b_1 = -1.50$, $b_2 = -1.00$, $b_3 = -0.50$, $b_4 = 0.00$ and $b_5 = 0.50$.

curves for the alternative information functions, we find the item information function, which is drawn by a thicker, dashed line. As was pointed out earlier, in the logistic model, all the six alternative information functions are positive for the entire range of θ , while the same is not true for the five alternative information functions in Model B. This result was expected from the result for the basic functions, which were observed earlier in this section.

The usefulness of the item information function has been emphasized earlier (Samejima, 1977), especially in connection with the maximum likelihood estimation of the examinee's ability. It should be noted, however, that the blind use of the item information function, or the test information function, is harmful, when the item response information functions, or the alternative information functions, are not always non-negative. This is exemplified in the criticism related with the three-parameter logistic model (Samejima, 1973). With models of high complexities, like Model B, care should be taken in finding out the limitation in using the item information function.

As the logical consequence of the observations made earlier for the basic functions, we find that for a certain interval of θ , which covers eight levels, the item information function in Model B is practically identical with the corresponding item information function in the logistic model. We can see in Figure 10-8-2 that this interval is approximately $(0.4, \infty)$. It is also noted that for this interval each alternative information function is practically identical with the counterpart in the logistic model, the fact which indicates that the affect of noises caused by random guessing is negligibly small in these intervals, and, therefore, we can expect that the accuracy of ability estimation is just as high as the one in the logistic model in these intervals of θ .

(X.9) Instructions and Mathematical Models

As was pointed out in Section X.5 , an interesting aspect of the role of mathematical models in psychology is that we have a control over stimuli to which persons react. In testing, not only can we develop the kinds of test items which, with an appropriate theory, enable us to measure the examinee's ability accurately, but also we can direct the examinee to react certain ways, by giving him suitable instructions.

If, for instance, the examinee is encouraged to guess randomly when he does not know the answer, then models like the three-parameter logistic model will be appropriate. It is not wise to give such instructions, however, since, in so doing, we fail to obtain and make use of the information given by the distractors. If our instructions extremely discourage the examinee to guess and convince him that it is wise to leave the question unanswered rather than to guess, then models like Bock's model and normal ogive model on the graded response level will be appropriate. If our instructions discourage the examinee to guess but encourage him to answer each question one way or another, then models like Models A, B and C will be appropriate.

There is no question that it is better to create the situation in which no noise is involved, and, therefore, we can use models like Bock's model or normal ogive model on the graded response level. It may not be easy to make suitable instructions for this purpose, however, without more or less deceiving the examinee. Besides, the pressure for success is so strong in most testing situations that the examinee may turn to guessing as the last resort regardless of the instructions. For this reason, Models A, B and C may be more realistic.

In any case, before selecting a specific model for our data, it is advisable to estimate the operating characteristics without assuming any mathematical form, using our methods and approaches which were introduced in Chapters 3, 5 and 6 , at least, for some test items. In so doing, if we find a substantially large number of examinees who skipped a test item, we shall estimate the operating characteristic for the "omission" of that item. If the estimated operating characteristic turned out to be informative, by providing us with a strictly decreasing function of ability or a unimodal function, then we decide that the category of "omission" is informative and use its operating characteristic in our process of ability estimation. If it turned out to be non-informative, by giving us no simple statistical relationship with ability or a constant function, then we decide the category of "omission" is useless in our process of ability estimation, and just ignore it.

(X.10) A New Approach to Data Analysis

Tables 10-10-1 and 10-10-2 present two contingency tables, each of which is for the four alternatives, A, B, C and D, of a multiple-choice test item and five ability groups of examinees. They were sampled from those made in the preliminary research conducted at Educational Testing Service, which had been given to the author by the courtesy of Mr. Donald Raske.

We notice in Table 10-10-1 that for this multiple-choice test item, Item 43, the mode of the frequency for the alternative A is the lowest ability group, that for the alternative B is the second highest ability group, that for C , the correct answer, is the highest ability group, and that for D is the lowest ability group. Thus these four alternatives may be arranged as A, D, B and C in ascending order. Actually these five ability groups are categorized with respect to the total test score and they may not give us very accurate categorization, and yet the table is informative enough to indicate the existence of differential information given by the four alternatives.

FIGURE 10-10-1

Contingency Table Between the Four Alternatives and the Five Ability Groups for Item 43 . (ETS Data)

Alternative	Very Low	Low	Middle	High	Very High	Total
A	55	40	26	26	12	159
B	64	58	68	80	52	322
C	34	64	70	74	130	372
D	34	36	29	19	6	124
No Answer	1	0	0	0	0	1
Total	188	198	193	199	200	978

FIGURE 10-10-2

Contingency Table Between the Four Alternatives and the Five Ability Groups for Item 46 . (ETS Data)

Alternative	Very Low	Low	Middle	High	Very High	Total
A	70	104	89	69	67	399
B	31	28	54	99	121	333
C	42	31	28	15	7	123
D	43	32	20	11	4	110
No Answer	0	2	0	0	0	2
Total	186	197	191	194	199	967

For Item 46, whose contingency table is given as Table 10-10-2, the modes are the second lowest ability group for the alternative A, the highest ability group for the alternative B, the correct answer, and the lowest ability group for both the alternatives C and D . Unlike Item 43, this test item does not have very explicit order for its four alternatives, since the alternatives C and D have similar frequency distributions. Thus the existence of differential information among the separate alternatives is less clear, and it may be advisable to examine the contents of the alternatives and replace some of them so that we shall obtain a contingency table similar to the one for Item 43 .

This type of contingency table is useful when we design our research project. It is advisable to select a set of test items which have confirmed content validity, and use its test score as the substitute for ability. In our preliminary study, we can make full use of the contingency table to eventually obtain a set of alternatives which provide us with differential information, so that we shall be able to adopt a model which belongs to the Informative Distractor Model (cf. Section IX.9).

It is expected in our contingency table that the correct answer shows strictly increasing frequencies. It is desirable that the other alternatives have differential modes among the four ability groups, "very low" through "high". It should be noted that we need one alternative for each test item which attracts examinees whose ability is low, in order to avoid the effect of random guessing for a wider interval of ability. This means that, in our contingency table, one alternative should have a distinctly high frequency for the lowest ability group.

As was discussed in Section VIII.6, if we find a subset, or subsets, of equivalent test items in the core set of test items of confirmed content validity, then we shall be able to apply Constant Information Model, and use the subset, or subsets, as Old Test. In such a case, the test items do not have to be equivalent in the sense that they have identical sets of item characteristic functions plus plausibility functions for all the wrong answers. What we need is the identical item characteristic functions, since the test items are treated as binary items when they are used as the substitute for the Old Test.

If we do not find such a subset of the core set of test items, then we shall assume a certain model for the correct answer of each test item, and estimate its parameters. In selecting the model, the contingency table for each core test item should be in consideration. Models like normal ogive model may serve for the purpose.

Now we shall estimate the maximum likelihood estimate of each examinee's ability, using our Old Test. Then we shall estimate the item characteristic function of each core test item, using one of the combinations of a method and an approach for the estimation, which were introduced in Chapters 3, 5 and 6 . This process is for the check of internal consistency, and, if the resultant estimated item characteristic function is close enough to the assumed one, we shall proceed. The plausibility function for each wrong answer of each core test item will be estimated, using our Old Test. Then we shall estimate the operating characteristic of each alternative of the test items, which are not included by the core set of test items, using the same method. After this has been accomplished, then we shall examine the resultant set of operating characteristics for each test item, and select an appropriate model. Note that we need not to choose a common model for all the test items. If we select several different models for separate subsets of our test items, we shall still be able to estimate the examinee's ability by the maximum likelihood estimation, provided that these models satisfy the unique maximum condition (cf. Samejima, 1969, 1972).

The above is a brief description of the new approach to data analysis. It is important that we select the core set of test items which have confirmed content validity. Psychometricians should not forget psychological reality, and the operational definition of ability dimension is by far the most important. In so doing, statistical techniques like factor analysis may be helpful in determining the dimensionality of ability. After this operational definition of ability has been accomplished with respect to the core set of test items, the operating characteristics of all the other test items must be estimated on this ability. Note that this new approach incorporates most of the main products of the present research, including the methods and approaches for estimating the operating characteristics of discrete item responses, the new family of models for the multiple-choice test item, Constant Information Model, and so forth.

REFERENCES

[1] Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick; Statistical theories of mental test scores. Addison-Wesley, 1986, Chapters 17-20.

[2] Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29-51.

[3] Lawley, D. N. and A.E. Maxwell. Factor analysis as a statistical method. London: Butterworth, 1971.

[4] Samejima. F. Application of the graded response model to the nominal response and multiple-choice situations. Chapel Hill, N.C.: University of North Carolina Psychometric Laboratory Report, 63, 1968.

[5] Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, No. 17, 1969.

[6] Samejima, F. A general model for free-response data. Psychometrika Monograph, No. 18, 1972.

[7] Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent trait theory. Psychometrika, 1973, 38, 221-233.

[8] Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, 233-247.

[9] Shiba, S. Construction of a scale for acquisition of word meanings. Bulletin of Faculty of Education, University of Tokyo, 17, 1968. (in Japanese)

XI Conclusions

The author has tried to integrats main topics and contents of the research within the preceding ten chapters of this final report. The work was difficult because of the abundance of products, and the author was forced to drop some relatively minor topics and contents. Because of the shortage of space, only a few figures and tables were selected for each chapter. To give an example, the estimated item characteristic functions are illutrated for item 6 only, and those for all the other nine binary test items are not shown in this final report. In spite of this fact, the author hopes that the reader will be assisted by this final report to increase his or her understanding of the contents of the research and its implications.

Estimation of the operating characteristics of discrete item responses, as well as that of ability distributions, without assuming any mathematical forms and using a relatively small number of examinees turned out to be successful. The fact that Subtest 6 with only eleven test items of three item score categories each proved to be sufficient to serve as Old Test in the estimation of the operating characteristics indicates the robustness of the present methods and approaches. We may be allowed to conclude this is a remarkable success. This finding may hold not only with unknown, binary test items, but also with unknown, graded test items and multiple-choice test items, or may not; the conclusion is yet to come. It is the author's wish that other researchers use the methods and approaches for different data to find out how they work.

The new family of models for the multiple-choice test item has been proposed and shown promise for the usefulness as models which belong to the Informative Distractor Model, to increase efficiency in ability estimation, and make the multiple-choice test item much more than a blurred image of the free-response test item. This family of models, combined with the methods and approaches for estimating the operating characteristics of discrete item responses, has given a new direction to research in mental measurement. The new approach must be tested in the near future upon empirical data, such as Shiba's. It is the author's hope that otheı researchers will also develop suitable test items and conduct research using the procedure proposed in this final report to find out how it works.

The method of moments for approximating any function by a polynomial, which proved to be the least squares solution, has effectively been incorporated. Constant Information Model has been proposed, and it has found its place in the new direction of research in mental measurement. Alternative estimators for the maximum likelihood estimator, which are population-free unlike Bayesian estimators, have been proposed, and it has been observed how they enhance the range of ability for which a specified test is effective and meaningful, keeping an approximate conditional unbiasedness, given ability.

The present research has produced many new theories and methods, and so forth. It has also proposed new problems and topics to pursue in the future. In this sense even though this is the final report of the present research, we are still in the middle of the way to advance latent trait theory and science.

DISTRIBUTION LIST

Navy

1 Dr. Alvah Bittner
Naval Biodynamics Laboratory
New Orleans, Lousiana 70189

1 Dr. Jack R. Borsting
Provost & Academic Dean
U.S. Naval Postgraduate School
Monterey, CA 93940

1 Dr. Robert Breaux
Code N-711
NAVTRAEQUIPCEN
Orlando, FL 32813

1 COMNAVMILPERSCOM (N-6C)
Dept. of Navy
Washington, DC 20370

1 CDR Mike Curran
Office of Naval Research
800 N. Quincy St.
Code 270
Arlington, VA 22217

1 Dr. Ric ard Elster
Department of Administrative Sciences
Naval Postgraduate School
Monterey, CA 93940

1 DR. PAT FEDERICO
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

1 Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. John Ford
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Henry M. Halff
Department of Psychology, C-009
University of California at San Diego
La Jolla, CA 92093

Navy

1 Dr. Patrick R. Harrison
Psychology Course Director
LEADERSHIP & LAW DEPT. (7b)
DIV. OF PROFESSIONAL DEVELOPMMENT
U.S. NAVAL ACADEMY
ANNAPOLIS, MD 21402

1 Dr. Norman J. Kerr
Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054

1 Dr. William L. Maloy
Principal Civilian Advisor for
Education and Training
Naval Training Command, Code 00A
Pensacola, FL 32508

1 Dr. Kneale Marshall
Scientific Advisor to DCNO(MPT)
OP01T
Washington DC 20370

1 CAPT Richard L. Martin, USN
Prospective Commanding Officer
USS Carl Vinson (CVN-70)
Newport News Shipbuilding and Drydock Co
Newport News, VA 23607

1 Dr. James McBride
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. George Moeller
Head, Human Factors Dept.
Naval Submarine Medical Research Lab
Groton, CN 06340

1 Ted M. I. Yellen
Technical Information Office, Code 201
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

1 Library, Code P201L
Navy Personnel R&D Center
San Diego, CA 92152

Navy

1 Technical Director
Navy Personnel R&D Center
San Diego, CA 92152

6 Commanding Officer
Naval Research Laboratory
Code 2627
Washington, DC 20390

1 Psychologist
ONR Branch Office
Bldg 114, Section D
666 Summer Street
Boston, MA 02210

1 Psychologist
ONR Branch Office
536 S. Clark Street
Chicago, IL 60605

1 Office of Naval Research
Code 437
800 N. Quincy SStreet
Arlington, VA 22217

5 Personnel & Training Research Programs
(Code 458)
Office of Naval Research
Arlington, VA 22217

1 Psychologist
ONR Branch Office
1030 East Green Street
Pasadena, CA 91101

1 Office of the Chief of Naval Operations
Research Development & Studies Branch
(OP-115)
Washington, DC 20350

1 Dr. Donald F. Parker
Graduate School of Business Administrati
University of Michigan
Ann Arbor, MI 48109

Navy

1 LT Frank C. Petho, MSC, USN (Ph.D)
Selection and Training Research Division
Human Performance Sciences Dept.
Naval Aerospace Medical Research Laborat
Pensacola, FL 32508

1 Director, Research & Analysis Division
Plans and Policy Department
Navy Recruiting Command
4015 Wilson Boulevard
Arlington, VA 22203

1 Dr. Bernard Rimland (03B)
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Worth Scanland, Director
Research, Development, Test & Evaluation
N-5
Naval Education and Training Command
NAS, Pensacola, FL 32508

1 Dr. Robert G. Smith
Office of Chief of Naval Operations
OP-987H
Washington, DC 20350

1 Dr. Alfred F. Smode
Training Analysis & Evaluation Group
(TAEG)
Dept. of the Navy
Orlando, FL 32813

1 Dr. Richard Sorensen
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Ronald Weitzman
Code 54 WZ
Department of Administrative Sciences
U. S. Naval Postgraduate School
Monterey, CA 93940

1 Dr. Robert Wisher
Code 309
Navy Personnel R&D Center
San Diego, CA 92152

Navy

1 DR. MARTIN F. WISKOFF
NAVY PERSONNEL R& D CENTER
SAN DIEGO, CA 92152

1 Mr John H. Wolfe
Code P310
U. S. Navy Personnel Research and
Development Center
San Diego, CA 92152

Army

1 Technical Director
U. S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Myron Fischl
U.S. Army Research Institute for the
Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Dexter Fletcher
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria,VA 22333

1 Dr. Milton S. Katz
Training Technical Area
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Harold F. O'Neil, Jr.
Attn: PERI-OK
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 LTC Michael Plummer
Chief, Leadership & Organizational
Effectiveness Division
Office of the Deputy Chief of Staff
for Personnel
Dept. of the Army
Pentagon, Washington DC 20301

1 DR. JAMES L. RANEY
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Mr. Robert Ross
U.S. Army Research Institute for the
Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

Army

1 Dr. Robert Sasmor
U. S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Commandant
US Army Institute of Administration
Attn: Dr. Sherrill
FT Benjamin Harrison, IN 46256

1 Dr. Frederick Steinheiser
Dept. of Navy
Chief of Naval Operations
OP-113
Washington, DC 20350

1 Dr. Joseph Ward
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Air Force

1 Air Force Human Resources Lab
AFHRL/MPD
Brooks AFB, TX 78235

1 U.S. Air Force Office of Scientific
Research
Life Sciences Directorate, NL
Bolling Air Force Base
Washington, DC 20332

1 Dr. Earl A. Alluisi
HQ, AFHRL (AFSC)
Brooks AFB, TX 78235

1 Research and Measurment Division
Research Branch, AFMPC/MPCYPR
Randolph AFB, TX 78148

1 Dr. Malcolm Ree
AFHRL/MP
Brooks AFB, TX 78235

1 Dr. Marty Rockway
Technical Director
AFHRL(OT)
Williams AFB, AZ 58224

1 Dr. Frank Schufletowski
U.S. Air Force
ATC/XPTD
Randolph AFB, TX 78148

Marines

1 H. William Greenup
Education Advisor (E031)
Education Center, MCDEC
Quantico, VA 22134

1 Major Howard Langdon
Headquarters, Marine Corps
OTTI 31
Arlington Annex
Columbia Pike at Arlington Ridge Rd.
Arlington, VA 20380

1 Director, Office of Manpower Utilization
HQ, Marine Corps (MPU)
BCB, Bldg. 2009
Quantico, VA 22134

1 Headquarters, U. S. Marine Corps
Code MPI-20
Washington, DC 20380

1 Special Assistant for Marine
Corps Matters
Code 100M
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

1 Major Michael L. Patrow, USMC
Headquarters, Marine Corps
(Code MPI-20)
Washington, DC 20380

1 DR. A.L. SLAFKOSKY
SCIENTIFIC ADVISOR (CODE RD-1)
HQ, U.S. MARINE CORPS
WASHINGTON, DC 20380

CoastGuard

1 Chief, Psychological Reserch Branch
U. S. Coast Guard (G-P-1/2/TP42)
Washington, DC 20593

1 Mr. Thomas A. Warm
U. S. Coast Guard Institute
P. O. Substation 18
Oklahoma City, OK 73169

Other DoD

12 Defense Technical Information Center
Cameron Station, Bldg 5
Alexandria, VA 22314
Attn: TC

1 Dr. William Graham
Testing Directorate
MEPCOM/MEPCT-P
Ft. Sheridan, IL 60037

1 Director, Research and Data
OASD(MRA&L)
3B919, The Pentagon
Washington, DC 20301

1 Military Assistant for Training and
Personnel Technology
Office of the Under Secretary of Defense
for Research & Engineering
Room 3D129, The Pentagon
Washington, DC 20301

1 Dr. Wayne Sellman
Office of the Assistant Secretary
of Defense (MRA & L)
2B269 The Pentagon
Washington, DC 20301

1 DARPA
1400 Wilson Blvd.
Arlington, VA 22209

Civil Govt

1 Dr. Lorraine D. Eyde
Personnel R&D Center
Office of Personnel Management of USA
1900 EStreet NW
Washington, D.C. 20415

1 Jerry Lehnus
REGIONAL PSYCHOLOGIST
U.S. Office of Personnel Management
230 S. DEARBORN STREET
CHICAGO, IL 60604

1 Dr. Andrew R. Molnar
Science Education Dev.
and Research
National Science Foundation
Washington, DC 20550

1 Dr. H. Wallace Sinaiko
Program Director
Manpower Research and Advisory Services
Smithsonian Institution
801 North Pitt Street
Alexandria, VA 22314

1 Dr. Vern W. Urry
Personnel R&D Center
Office of Personnel Management
1900 E Street NW
Washington, DC 20415

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

Non Govt

1 Dr. Erling B. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

1 1 psychological research unit
Dept. of Defense (Army Office)
Campbell Park Offices
Canberra ACT 2600, Australia

1 Dr. Isaac Bejar
Educational Testing Service
Princeton, NJ 08450

1 Capt. J. Jean Belanger
Training Development Division
Canadian Forces Training System
CFTSHQ, CFB Trenton
Astra, Ontario KOK 1B0

1 Dr. John Bergan
School of Education
University of Arizona
Tuscon AZ 85721

1 CDR Robert J. Biersner
Program Manager
Human Performance
Navy Medical R&D Command
Bethesda, MD 20014

1 Dr. Menucha Birenbaum
School of Education
Tel Aviv University
Tel 4viv, Ramat Aviv 69978
Israel

1 Dr. Werner Birke
DezWPs im Streitkraefteamt
Postfach 20 50 03
D-5300 Bonn 2
WEST GERMANY

1 Dr. R. Darrel Bock
Department of Education
University of Chicago
Chicago, IL 60637

Non Govt

1 Liaison Scientists
Office of Naval Research,
Branch Office , London
Box 39 FPO New York 09510

1 Col Ray Bowles
800 N. Quincy St.
Room 804
Arlington, VA 22217

1 Dr. Robert Brennan
American College Testing Programs
P. O. Box 168
Iowa City, IA 52240

1 DR. C. VICTOR BUNDERSON
WICAT INC.
UNIVERSITY PLAZA, SUITE 10
1160 SO. STATE ST.
OREM, UT 84057

1 Dr. Anthony Cancelli
School of Education
University of Arizona
Tuscon, AZ 85721

1 Dr. John B. Carroll
Psychometric Lab
Univ. of No. Carolina
Davie Hall 013A
Chapel Hill, NC 27514

1 Charles Myers Library
Livingstone House
Livingstone Road
Stratford
London E15 2LJ
ENGLAND

1 Dr. Kenneth E. Clark
College of Arts & Sciences
University of Rochester
River Campus Station
Rochester, NY 14627

Non Govt

1 Dr. Norman Cliff
Dept. of Psychology
Univ. of So California
University Park
Los Angeles, CA 90007

1 Dr. William E. Coffman
Director, Iowa Testing Programs
334 Lindquist Center
University of Iowa
Iowa City, IA 52242

1 Dr. Meredith P. Crawford
American Psychological Association
1200 17th Street, N.W.
Washington, DC 20036

1 Director
Behavioural Sciences Division
Defence & Civil Institute of
Environmental Medicine
Post Office Box 2000
Downsview, Ontario M3M 3B9
CANADA

1 Dr..Fritz Drasgow
Yale School of Organization and Manageme
Yale University
Box 1A
New Haven, CT 06520

1 Dr. Mavin D. Dunnette
Personnel Decisions Research Institute
2415 Foshay Tower
821 Marguette Avenue
Mineapolis, MN 55402

1 Mike Durmeyer
Instructional Program Development
Building 90
NET-PDCD
Great Lakes NTC, IL 60088

1 ERIC Facility-Acquisitions
4833 Rugby Avenue
Bethesda, MD 20014

Non Govt

1 Dr. Benjamin A. Fairbank, Jr.
McFann-Gray & Associates, Inc.
5825 Callaghan
Suite 225
San Antonio, Texas 78228

1 Dr. Leonard Feldt
Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 Dr. Victor Fields
Dept. of Psychology
Montgomery College
Rockville, MD 20850

1 Univ. Prof. Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

1 Professor Donald Fitzgerald
University of New England
Armidale, New South Wales 2351
AUSTRALIA

1 Dr. John R. Frederiksen
Bolt Beranek & Newman
50 Moulton Street
Cambridge, MA 02138

1 DR. ROBERT GLASER
LRDC
UNIVERSITY OF PITTSBURGH
3939 O'HARA STREET
PITTSBURGH, PA 15213

1 Dr. Ron Hambleton
School of Education
University of Massachusetts
Amherst, MA 01002

Non Govt

1 Dr. Chester Harris
School of Education
University of California
Santa Barbara, CA 93106

1 Dr. Lloyd Humphreys
Department of Psychology
University of Illinois
Champaign, IL 61820

1 Library
HumRRO/Western Division
27857 Berwick Drive
Carmel, CA 93921

1 Dr. Steven Hunka
Department of Education
University of Alberta
Edmonton, Alberta
CANADA

1 Dr. Jack Hunter
2122 Coolidge St.
Lansing, MI 48906

1 Dr. Huynh Huynh
College of Education
University of South Carolina
Columbia, SC 29208

1 Professor John A. Keats
University of Newcastle
AUSTRALIA 2308

1 Dr. Michael Levine
Department of Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

1 Dr. Charles Lewis
Faculteit Sociale Wetenschappen
Rijksuniversiteit Groningen
Oude Boteringestraat 23
9712GC Groningen
Netherlands

Non Govt

1 Dr. Robert Linn
College of Education
University of Illinois
Urbana, IL 61801

1 Dr. Frederick M. Lord
Educational Testing Service
Princeton, NJ 08540

1 Dr. James Lumsden
Department of Psychology
University of Western Australia
Nedlands W.A. 6009
AUSTRALIA

1 Dr. Gary Marco
Educational Testing Service
Princeton, NJ 08450

1 Dr. Scott Maxwell
Department of Psychology
University of Houston
Houston, TX 77004

1 Dr. Samuel T. Mayo
Loyola University of Chicago
820 North Michigan Avenue
Chicago, IL 60611

1 Professor Jason Millman
Department of Education
Stone Hall
Cornell University
Ithaca, NY 14853

1 Dr. Melvin R. Novick
356 Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. Jesse Orlansky
Institute for Defense Analyses
400 Army Navy Drive
Arlington, VA 22202

1 Dr. James A. Paulson
Portland State University
P.O. Box 751
Portland, OR 97207

Non Govt

1 MR. LUIGI PETRULLO
2431 N. EDGEWOOD STREET
ARLINGTON, VA 22207

1 DR. DIANE M. RAMSEY-KLEE
R-K RESEARCH & SYSTEM DESIGN
3947 RIDGEMONT DRIVE
MALIBU, CA 90265

1 MINRAT M. L. RAUCH
P II 4
BUNDESMINISTERIUM DER VERTEIDIGUNG
POSTFACH 1328
D-53 BONN 1, GERMANY

1 Dr. Mark D. Reckase
Educational Psychology Dept.
University of Missouri-Columbia
4 Hill Hall
Columbia, MO 65211

1 Dr. Leonard L. Rosenbaum, Chairman
Department of Psychology
Montgomery College
Rockville, MD 20850

1 Dr. Lawrence Rudner
403 Elm Avenue
Takoma Park, MD 20012

1 Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

1 DR. ROBERT J. SEIDEL
INSTRUCTIONAL TECHNOLOGY GROUP
HUMRRO
300 N. WASHINGTON ST.
ALEXANDRIA, VA 22314

1 Dr. Kazuo Shigemasu
University of Tohoku
Department of Educational Psychology
Kawauchi, Sendai 980
JAPAN

Non Govt

1 Dr. Edwin Shirkey
Department of Psychology
University of Central Florida
Orlando, FL 32816

1 Dr. Robert Smith
Department of Computer Science
Rutgers University
New Brunswick, NJ 08903

1 Dr. Richard Snow
School of Education
Stanford University
Stanford, CA 94305

1 DR. PATRICK SUPPES
INSTITUTE FOR MATHEMATICAL STUDIES IN
THE SOCIAL SCIENCES
STANFORD UNIVERSITY
STANFORD, CA 94305

1 Dr. Hariharan Swaminathan
Laboratory of Psychometric and
Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

1 Dr. Brad Sympson
Psychometric Research Group
Educational Testing Service
Princeton, NJ 08541

1 Dr. Kikumi Tatsuoka
Computer Based Education Research
Laboratory
252 Engineering Research Laboratory
University of Illinois
Urbana, IL 61801

1 Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

1 Dr. Robert Tsutakawa
Department of Statistics
University of Missouri
Columbia, MO 65201

Non Govt

1 Dr. James Chen
354 Lindquist Center for Measurement
University of Iowa
Iowa City, Iowa 52240

1 Dr. Douglas Carroll
Bell Laboratories
600 Mountain Ave.
Murray Hill, N.J. 07974

1 Dr. Robert Guion
Department of Psychology
Bowling Green State University
Bowling Green, OH 43403

1 Dr. Niel Timm
Department of Information Systems
Planning
University of Pittsburgh
Pittsburgh, PA 15260

1 Dr. Bert F. Green
Department of Psychology
The John's Hopkins University
Charles at 34th Street
Baltimore, MD 21218

1 Dr. Ron Hambleton
School of Education
University of Massachusetts
Amherst, MA 01002

1 Dr. William W. Turnbull
Educational Testing Service
Princeton, NJ 08540

1 Dr. Isaac I. Bejar
Department of Psychology
Elliot Hall
75 East River Road
Minneapolis, MN 55455

1 Dr. Lowell Schipper
Department of Psychology
Bowling Green State University
Bowling Green, OH 43403

Non Govt

1 Dr. P. Mengal
Faculte' de Psychologie
et des Sciences de l'Education
Universite' de Geneve
3 fl. de l'Universite
1201 Geneva SWITZERLAND

1 Dr. Wim J. van der Linden
Vakgroep Onderwijskunde
Postbus 217
7500 EA Enschede
The Netherlands

1 Dr. Lutz Hornke
University Duesseldorf
Erz. Wiss.
D-4000 Duesseldorf
WEST GERMANY

1 Dr. Wolfgang Buchtala
8346 Simbach Inn
Postfach 1306
Industriestrasse 1
WEST GERMANY

1 Dr. Albert Beaton
Educational Testing Service
Princeton, New Jersey 08450

1 Dr. Sukeyori Shiba
Faculty of Education
University of Tokyo
Hongo, Bunkyoku
Tokyo, Japan 113

1 Mr. Yukihiro Noguchi
Faculty of Education
University of Tokyo
Hongo, Bunkyoku
Tokyo, Japan 113

1 Dr. Takahiro Sato (Representative)
Application Research Laboratory
Central Research Laboratories
Nippon Electric Co., Ltd.
4-1-1 Miyazaki, Takatsu-ku
Kawasaki 213, Japan

Non Govt

1 Dr. J. Uhlaner
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91364

1 Dr. Howard Wainer
Division of Psychological Studies
Educational Testing Service
Princeton, NJ 08540

1 Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455

1 DR. SUSAN E. WHITELY
PSYCHOLOGY DEPARTMENT
UNIVERSITY OF KANSAS
LAWRENCE, KANSAS 66044

1 Wolfgang Wildgrube
Streitkraefteamt
Box 20 50 03
D-5300 Bonn 2

Navy

1 Ms. Nancy McMan
Office of Naval Research
206 O'Keefe Building
Atlanta, GA 30332

Army

1 Dr. Randall M. Chambers
U.S. Army Research Institute
for the Behavioral & Social
Sciences
Fort Sill Field Unit
P.O. Box 3066
Fort Sill, OK 73503